

ENTROPY AND OPTIMIZATION

A thesis submitted in partial fulfillment
of the requirements for the degree of

Doctor of Philosophy

The University of Liverpool

1987

LI XING-SI

ABSTRACT

The present research explores the use of Shannon's (informational) entropy measure and Jaynes' maximum entropy criterion in the solution of constrained optimization problems. At present the methods of mathematical programming contain no information-theoretic basis. They all view the optimization problems in terms of a topological domain defined deterministically by function hypersurfaces. Information theory appears to be incompatible with this as it is essentially concerned with probabilities.

In contrast with the existing optimization methods, we simulate an optimization problem as a statistical thermodynamic system. Thus an optimizing process may then be interpreted as transitions of the system to a sequence of equilibrium states which are characterized by certain entropy maxima depending upon a "temperature" parameter.

Several ways of using entropy in the optimization context are investigated. Some entropy-based methods are developed. In particular a primal function for constrained optimization is derived which has some of the characteristics of the norm of a vector and may be a basis for developing new and radically different algorithms. The present approach is similar but superior to the least pth method for minimax optimization.

A class of structural optimization problems is chosen to provide example problems. Computational results demonstrate that the entropy-based algorithm proposed in this work is efficient and reliable.

The following parts of the present work are original:

1. Two ways of deriving the dual of the maximum entropy problem are presented, the first using classical Lagrangean saddle-point criteria, and the second using the well-established duality theory of geometric programming. These dual forms yield additional insight into the nature of entropic processes and clarify some of ambiguities in the literature. They also afford a simple means of calculating the least biased probabilities using only standard unconstrained minimization algorithms.
2. The framework of a surrogate constraint approach is utilized to simulate an optimization problem as a statistical thermodynamic system. The system approaches spontaneously an equilibrium state which is characterized by a certain maximum entropy of the system associated with a specified "temperature". Therefore an entropy maximization problem is formulated to estimate probabilities, i.e., surrogate multipliers, of the system being in each micro-state. Another approach to estimating these multipliers is also proposed which adds a term related to the multiplier entropy to a surrogate dual. Then a dual iterative procedure is derived using Lagrangean stationarity conditions for the modified surrogate dual problem. It is proved that the entropy-augmented dual approach is equivalent to the entropy maximization approach.

3. The entropy-augmented approach is further developed to derive a primal function for general constrained optimization which has some of the characteristics of the pth norm of a vector and approaches the uniform norm as the parameter p tends to infinity. It is also proved by means of the arithmetic-geometric inequality that this function actually produces an upper bound on the minimax Lagrangean. From a computational viewpoint, this primal function may be a basis for developing new and radically different algorithms.
4. An explicit dual for a class of structural optimization problems is derived through the surrogate approach and it turns out to be the same as that by Templeman[91]. An entropy-based algorithm is presented to solve the optimum design problem. A remarkable feature of the proposed algorithm is that it retains all of the original constraints in the optimization procedure and increases the computational efforts only marginally. The iterations involve only algebraic operations so that the algorithm is very simple and easy to program. The results are compared with those obtained by other methods and show that the present algorithm is very effective and reliable.

ACKNOWLEDGEMENTS

The author wishes to express his deep gratitude and sincere thanks to his supervisor, Dr. A.B.Templeman, for his encouragement in initiating the present research and generous hours of discussion which the author received during the course of the present work, and for his help in finding the necessary financial support.

The author wishes also to acknowledge gratefully the financial support provided by both the British Government under the ORS scheme and The University of Liverpool for granting a postgraduate studentship.

He is also indebted to the Chinese Government for additional financial support and to Dalian Institute of Techonology for giving leave to pursue this study.

NOTATIONS

c^1 - continuous and once differentiable functions

$E[.]$ - expectation operator

E^n - n-dimensional Euclidean space

exp - the exponent function

GP - geometric programming

\lim - the limit

\ln - the natural logarithmic function

LP - linear programming

Max - maximize

MEC - Maximum Entropy Criterion

MED - Maximum Entropy (probability) Distribution

Min - minimize

NLP - nonlinear programming

p.d. - probability distribution

r.v. - random variable

s.t. - subject to

w.r.t. - with respect to

∇f - gradient of the function f

CONTENTS

ABSTRACT	i
ACKNOWLEDGEMENTS	iv
NOTATIONS	v
Chapter 1 Introduction	1
Summary	1
1.1 Introduction	1
1.2 Constrained Optimization Theory	4
1.2.1 Convexity	5
1.2.2 Lagrange Multipliers and Kuhn-Tucker Conditions	8
1.2.3 Saddle Point and Duality	13
1.3 Constrained Optimization Methods	16
1.3.1 Primal Methods	18
1.3.2 Penalty Function Methods	22
1.3.3 Lagrange Methods	25
1.3.4 Multiplier Methods	26
1.4 Surrogate Approach to Constrained Optimization	28
1.4.1 Some Relationships between Problems (P) and (S)	29
1.4.2 Surrogate Solution Procedure	31
1.4.3 Surrogate Duality	33
1.5 Motivations and Specifications of the Present Research	35

1.5.1	Motivations of the Present Work	36
1.5.2	Specifications of the Present Work	38
Chapter 2	Entropy and The Maximum Entropy Criterion	44
Summary	44
2.1	Introduction	45
2.2	Definitions and Properties of Informational Entropy	48
2.3	The Maximum Entropy Criterion (MEC)	51
2.4	Interpretation of the Maximum Entropy Criterion	56
2.5	Some Applications of the Maximum Entropy Criterion	57
2.6	Applications of Entropy to Optimization Problems	60
2.7	Conclusions and Relevance to the Present Work	65
Chapter 3	Duals of the Maximum Entropy Problem	68
Summary	68
3.1	Introduction	68
3.2	Dual Forms of the Maximum Entropy Problem (MEC)	70
3.2.1	Lagrangean Dual Formulation	70
3.2.2	Geometric Programming Dual Formulation	71
3.3	Interpretation of the Entropy Duals	74
3.4	Computational Application of the Entropy Duals	76
3.5	Dual Forms for the Continuous Case	80
3.6	Conclusions	81
Chapter 4	Entropy in Optimization: Dual Methods	82
Summary	82
4.1	Introduction	83

4.2	Statistical Thermodynamic Simulation of Optimization	85
4.3	An Entropy Maximization Approach to Optimization . .	89
4.4	An Entropy Augmentation Approach to Optimization . .	94
4.5	Equivalence Relationship between Eqs.(4.11) and (4.31)	100
4.6	Conclusions	102
Chapter 5 Entropy in Optimization: Primal Methods . .		103
	Summary	103
5.1	Introduction	104
5.2	Primal Function of Constrained Optimization	106
5.3	Some Important Properties of the Primal Function .	111
5.4	Development of Numerical Algorithms	118
5.5	Discussions	122
5.6	Conclusions	125
Chapter 6 Entropy Methods for Structural Optimization		128
	Summary	128
6.1	Introduction	130
6.2	Primal-Dual Method for Structural Optimization . .	135
6.3	Primal-Dual Algorithm for Structural Optimization .	139
6.4	Numerical Examples	142
6.5	Surrogate Dual for Structural Optimization	145
	6.5.1 Lagrangean Dual Formulation	146
	6.5.2 Geometric Programming Dual Formulation	147
6.6	Entropy-Based Dual Method for Structural Optimization	150
6.7	Conclusions	154

Chapter 7	Conclusions and Recommendations for Future Work	163
7.1	Introduction	163
7.2	Conclusions	165
7.3	Recommendations for Future Work	167
REFERENCES	171
APPENDIX A	GEOMETRIC PROGRAMMING	181
APPENDIX B	PUBLISHED PAPERS	184
Appendix B1	Entropy Duals	185
Appendix B2	A Maximum Entropy Approach to Constrained Non-Linear Programming	199

CHAPTER 1 INTRODUCTION

SUMMARY

This chapter introduces some of the basic concepts and important results of constrained optimization theory that are frequently used in the literature and in this thesis. Also, there is a brief review of the existing methods for solving constrained optimization problems which are either computationally effective or theoretically important. In particular, the surrogate constraint approach is given special attention due to its close relevance to the present work.

The present research is motivated by a desire to provide optimization with an information-theoretic basis, and thereby to develop new and radically different optimization methods.

1.1 INTRODUCTION

The solution of an engineering decision-making problem invariably requires a compromise among many alternatives. The engineer wishes to seek an ideal compromise that satisfies all the limitations and

restrictions imposed on the decision problem and is best in some sense, such as maximum efficiency, minimum cost, or minimum weight, etc.

Mathematical optimization techniques together with powerful computers enable one to systematically modify the decision towards the "optimum". Using this approach, one formulates a complex decision problem as the following optimization problem:

$$\text{Min } f(x) \quad (1.1)$$

$$\text{s.t. } g_j(x) \leq 0 \quad j=1, \dots, m \quad (1.2)$$

$$\text{and/or } h_k(x) = 0 \quad k=1, \dots, \ell \quad (1.3)$$

where $x = (x_1, \dots, x_n)$ is a n -vector in E^n representing adjustable parameters, and $f(x)$, $g_j(x)$ and $h_k(x)$ are real-valued functions of the vector x .

The function $f(x)$ is called the objective function of the problem, and the inequalities (1.2) and equalities (1.3) are the constraints which represent the restrictions imposed on the vector x of adjustable parameters.

Any vector x that satisfies all the constraints (1.2) and (1.3) is called a feasible point. The set of all the feasible points constitutes the feasible region which is sometimes called constraint set.

A point x^* , which satisfies all the constraints and at which $f(x)$ attains its minimum, is called the optimum point, and the pair x^* and

$f(x^*)$ constitutes an optimum solution where $f(x^*)$ represents the optimum value of the objective $f(x)$. If $f(x^*) \leq f(x)$ holds for all x in the feasible region, then x^* is called global optimum point, otherwise it is a local optimum point.

An inequality constraint $g_j(x)$ is said to be active if $g_j(x^*) = 0$ at the optimum point x^* . For a real-world problem, there almost always exist several constraints, so that the corresponding problem is constrained, linearly or nonlinearly. Throughout this thesis, only constrained problems are considered.

An optimization problem is said to be a linear programming problem (LP) if all the problem functions are linear in the variables x_i , otherwise it is a nonlinear programming problem (NLP). The present work is mainly concerned with nonlinear programming problems.

The area of linear programming in the last three decades has achieved great success - the most significant development being the simplex method by which the solutions of most linear programming problems can be systematically obtained in a finite number of iterations. On the other hand, there has been no single algorithm devised that can handle all nonlinear problems. The major difficulties presented in solving nonlinear optimization problems arise from the presence of nonlinear constraints. Keeping them satisfied involves considerable complexity. When inequality constraints are present it is impossible for one to know in advance whether an inequality constraint will be active or not at

the optimum. This has left the area of nonlinear optimization in an unsatisfactory state and it is, therefore, an active research field.

One of the objectives of the present work is to look at nonlinear programming at a different angle from previous work and to develop new, maybe more efficient, algorithms.

The organization of this chapter is as follows. Section 1.2 introduces some important concepts and results of constrained optimization theory such as convexity, Lagrange multipliers, Kuhn-Tucker conditions, saddle point, duality, etc.; Section 1.3 gives a brief review of the existing methods for solving constrained optimization problems; Section 1.4 examines a special approach, the so-called surrogate approach, which has close links to the present research. Motivations and specifications of the present work conclude the chapter.

1.2 CONSTRAINED OPTIMIZATION THEORY

Several concepts and results such as convexity, Lagrange multipliers, Kuhn-Tucker conditions, saddle point and duality are closely related and are very important in the development of constrained optimization from both theoretical and computational viewpoints. They are introduced

below without proofs which can be found elsewhere (Refs.[34], [39], [68], [71]).

1.2.1 CONVEXITY

A set is said to be convex if, given any two points x_1 and x_2 in the set, the point $x=\lambda x_1+(1-\lambda)x_2$ is also in the set where λ is a scalar, $0\leq\lambda\leq 1$.

This definition can be interpreted geometrically as stating that a set is convex if, given any two points in the set, every point on the line segment joining these two points is also in the set. This is illustrated in Fig.1.1.

The following proposition shows that certain set operations preserve convexity.

Proposition 1.1: convex sets in E^n satisfy the following relations

1. If C is a convex set and β is a real number, the set βC is convex.
2. If C and D are convex, then the set $C+D$ is convex.
3. The intersection of any collection of convex sets is convex.

A function $f(x)$ is said to be a convex function if for all x_1, x_2 in the (convex) domain of definition of $f(x)$ and for all λ ($0 \leq \lambda \leq 1$), the following inequality

$$f[\lambda x_1 + (1-\lambda)x_2] \leq \lambda f(x_1) + (1-\lambda)f(x_2) \quad (1.4)$$

holds. A concave function $f(x)$ is defined in (1.4) if the relation " \leq " is replaced by " \geq ".

A linear function is both convex and concave. A positive linear combination of convex functions is also convex.

The definition of a convex function can also be interpreted geometrically. A function $f(x)$ is said to be convex if the line segment drawn between any two points on the graph of the function never lies below the graph, and concave if the line segment never lies above the graph. The convexity of functions is illustrated in Fig.1.2.

Occasionally, quasi-convexity is used. A function $f(x)$ is said to be quasi-convex in a convex set if for any two points x_1 and x_2 in the set, the following relation

$$f(x) \leq \max(f(x_1), f(x_2)) \quad (1.5)$$

holds, where the point x is on the line segment joining the given points, i.e., $x = \lambda x_1 + (1-\lambda)x_2$ ($0 \leq \lambda \leq 1$). A quasi-concave function $f(x)$ is defined if the relation (1.5) is replaced by $f(x) \geq \min(f(x_1), f(x_2))$.

A convex programming problem is one of minimizing a convex function (or maximizing a concave function) over a convex constraint set. For a convex programming problem, any local optimum is a global optimum. This is a desirable property from which many good methods are derived.

The following proposition together with Proposition 1.1 can help identify whether or not a constraint set is convex.

Proposition 1.2: If $g(x)$ is convex, then the set $R=\{x: g(x)\leq b\}$ is convex for all scalars b .

As a consequence of Propositions 1.1 and 1.2, the inequality constrained problem defined by (1.1) and (1.2), referred to as problem (P) later on, is a convex programming provided that the functions $f(x)$ and $g_j(x)$, $j=1,\dots,m$, are all convex. This is true because: (1) each of the sets $R_j=\{x: g_j(x)\leq 0\}$ is convex by Proposition 1.2, and (2) the constraint set R is the intersection of the sets R_j and is convex by Proposition 1.1(3).

Although convexity is desirable, many real-world problems turn out to be nonconvex. In addition, there is no simple way to test a nonlinear problem for convexity, since there is no simple way to test a nonlinear function for this property. However convexity plays an important role in the development of nonlinear optimization methods. When convexity is assumed, many significant results have been derived. Often these results can give insight into the properties of more general problems. Sometimes, such results may be carried over directly to nonconvex

problems. Here the convexity assumption has given the researcher a basis upon which to derive the method, and the method may then be applicable to more general situations. Convexity thus plays a role in optimization theory much the same as that of linearity in mechanics. It is well known that many results derived from linear theory are used in nonlinear analyses.

1.2.2 LAGRANGE MULTIPLIERS AND KUHN-TUCKER CONDITIONS

Lagrange multipliers and the Kuhn-Tucker conditions are closely related and are very important in constrained optimization from both theoretical and computational viewpoints since they provide means for dealing with the constraints and establish criteria for recognizing the optima.

Lagrange multipliers

The equality constrained problem

(E):

$$\text{Min} \quad f(x) \quad (1.1)$$

$$\text{s.t.} \quad h_k(x) = 0 \quad k=1, \dots, \ell \quad (1.3)$$

may be indirectly solved by constructing a Lagrangian function:

$$L(x, \mu) = f(x) + \sum_{k=1}^{\ell} \mu_k h_k(x) \quad (1.6)$$

where μ_k ($k=1, \dots, \ell$) are called the Lagrange multipliers.

The Lagrange multiplier of each constraint measures the rate of changes in the objective function, consequent upon changes in that constraint function. This information can be valuable in that it indicates how sensitive the objective function is to changes in the different constraints.

It can be proved that the solution of problem (E) corresponds to a stationary point of the Lagrangian (1.6). Necessary conditions for stationarity are:

$$\partial L(x, \mu) / \partial x_i = \partial f(x) / \partial x_i + \sum_{k=1}^{\ell} \mu_k (\partial h_k(x) / \partial x_i) = 0 \quad i=1, \dots, n \quad (1.7)$$

and

$$\partial L(x, \mu) / \partial \mu_k = h_k(x) = 0 \quad k=1, \dots, \ell \quad (1.8)$$

which reduce the problem (E) to solving $n+\ell$ simultaneous equations in the $n+\ell$ variables x_i ($i=1, \dots, n$) and μ_k ($k=1, \dots, \ell$).

However, it is not easy to solve the system of equations (1.7) and (1.8) for the following reasons. First of all, the problem

dimensionality is considerably increased since iterations have to be carried out simultaneously on x and μ . Moreover Eqs.(1.7) and (1.8) are usually coupled and nonlinear so that they are hard to solve analytically or numerically. Because of this, the Lagrange multiplier method is not very useful for solving constrained optimization problems.

It should also be noted that not all solutions of Eqs.(1.7) and (1.8) will be constrained minima, since these conditions also pertain to constrained maxima and saddle points.

Kuhn-Tucker Conditions

The most important theoretical results in the field of nonlinear optimization are the Kuhn-Tucker conditions[67] that establish the necessary conditions for optimality of the inequality constrained optimization problems. These conditions must be satisfied at any constrained optimum, local or global, of any linear and most nonlinear optimization problems. They form the basis for the development of many computational algorithms. In addition, some criteria for stopping many algorithms, i.e., for recognizing when a local constrained optimum has been achieved, are derived directly from them.

The geometrical and analytical statements of the Kuhn-Tucker conditions are described as follows.

The Geometrical Statement of the Kuhn-Tucker Conditions: if f and all g_j are differentiable, the necessary conditions for a point x^* to be a constrained minimum of the problem

(P)

$$\text{Min} \quad f(x) \quad (1.1)$$

$$\text{s.t.} \quad g_j(x) \leq 0 \quad j=1, \dots, m \quad (1.2)$$

are that at x^* , $-\nabla f$ lie within the cone generated by the gradients $\nabla g_j(x^*)$ of the active constraints.

The geometrical significance of these conditions is illustrated in Fig.1.3.

The Kuhn-Tucker conditions are closely related to the classical Lagrange multiplier results Eqs.(1.6)-(1.8). By analogy with problem (E), a Lagrangean function for problem (P) can be constructed as

$$L(x, \mu) = f(x) + \sum_{j=1}^m \mu_j g_j(x) \quad (1.9)$$

where μ_j are viewed as Lagrange multipliers for the inequality constraints $g_j(x) \leq 0$. The corresponding necessary conditions for optimality are analytically stated below.

The Analytical Statement of the Kuhn-Tucker Conditions: If x^* is a local minimizer of the problem (P) and is also a regular point, then

there exist Lagrange multipliers μ^* such that x^* , μ^* satisfy the following conditions:

$$\partial L(x^*, \mu^*) / \partial x_i = \partial f(x^*) / \partial x_i + \sum_{j=1}^m \mu_j^* (\partial g_j(x^*) / \partial x_i) = 0 \quad i=1, \dots, n \quad (1.10)$$

$$g_j(x^*) \leq 0 \quad j=1, \dots, m \quad (1.11)$$

$$\mu_j^* \geq 0 \quad j=1, \dots, m \quad (1.12)$$

$$\mu_j^* g_j(x^*) = 0 \quad j=1, \dots, m \quad (1.13)$$

where x^* is said to be a regular point of the constraint set if the gradient vectors of the active constraints at x^* are linearly independent[71].

A point x^* which satisfies the Kuhn-Tucker conditions is sometimes referred to as a K-T point. Eq.(1.13) is referred to as the complementarity condition.

Although the Kuhn-Tucker necessary conditions have great significance in constrained optimization theory, they cannot be used directly to solve optimization problems. The reasons are similar to those noted earlier for stationarity conditions (1.7) and (1.8) of the Lagrangean function (1.6).

1.2.3 SADDLE POINT AND DUALITY

Saddle Point

The Kuhn-Tucker conditions have a restriction on their application which is that the objective and constraint functions must be differentiable. There are other conditions, again centering around the Lagrangean function, which hold even in the absence of differentiability. These are the saddle-point criteria and are sufficient for a given point to be optimum for almost any optimization problem.

A point (x', μ') with $\mu' \geq 0$ is said to be a saddle point for the Lagrangean $L(x, \mu)$ shown as in (1.9) if it satisfies

$$L(x', \mu') \leq L(x, \mu') \quad \text{for all } x \quad (1.14)$$

$$L(x', \mu') \geq L(x', \mu) \quad \text{for all } \mu \geq 0 \quad (1.15)$$

That is, x' minimizes $L(x, \mu')$ over x -space and μ' maximizes $L(x', \mu)$ over all $\mu \geq 0$. In two variables, $L(x, \mu)$ would have the saddle shape with the saddle point yielding simultaneously a minimum in one variable and a maximum in the other.

The following theorem gives necessary and sufficient conditions for a saddle point of $L(x, \mu)$.

Saddle-Point Theorem: (x', μ') is a saddle point for $L(x, \mu)$ if and only if

$$(a) \ x' \text{ minimizes } L(x, \mu') \text{ over } x\text{-space} \quad (1.16)$$

$$(b) \ g_j(x') \leq 0 \quad j=1, \dots, m \quad (1.17)$$

$$(c) \ \mu' \geq 0 \quad j=1, \dots, m \quad (1.18)$$

$$(d) \ \mu_j' g_j(x') = 0 \quad j=1, \dots, m \quad (1.19)$$

Similarity of the saddle-point conditions (1.16)-(1.19) to the Kuhn-Tucker conditions (1.10)-(1.13) should be noted. Conditions (b)-(d) are common to both. Condition (a) replaces the stationarity condition (1.10) of $L(x, \mu)$ by its minimization. They will be proved to be equivalent for convex differentiable programming problems.

The usefulness of a saddle point is brought out from the fact that if (x', μ') is a saddle point for $L(x, \mu)$, then x' solves the primal problem (P).

Duality

The concept of duality occurs widely in the optimization literature. The aim is to provide an alternative formulation of an optimization problem which is more convenient computationally or has some theoretical significance. The original problem is referred to as the primal and the transformed problem as the dual. Usually the variables in the dual can

be interpreted as Lagrange multipliers and take the value μ^* at the dual solution, where μ^* is a multiplier vector associated with a primal solution x^* . In this sense the Lagrange multiplier methods might be thought of as a dual method. Usually, however, there is also present an objective function which has to be optimized. Duality theory of this kind is associated with a convex programming problem as the primal, and it is important to realize that if the primal problem is not convex then the dual problem may well not have a solution from which the primal solution can be derived.

A particular form of dual formulation was proposed by Wolfe[105]. It may be stated as

Wolfe's Dual Problem:

If x^* solves the convex primal problem and is a regular point, and if the objective and constraint functions are in C^1 , then x^* and μ^* solve the dual problem

$$\begin{array}{ll} \text{Max} & L(x, \mu) \\ \text{over} & x, \mu \end{array} \quad (1.20)$$

$$\text{s.t.} \quad \nabla_x L(x, \mu) = 0 \quad (1.21)$$

$$\mu \geq 0 \quad (1.22)$$

Furthermore the minimum primal and the maximum dual values are equal, that is, $f(x^*) = L(x^*, \mu^*)$.

The nonnegativity condition (1.22) of μ applies only to problem (P).

Wolfe's dual formulation requires the differentiability of the objective and constraint functions. An alternative dual formulation that does not have the requirement of differentiability and sometimes is convenient computationally was proposed by Falk[32]. Falk's dual formulation takes the following form:

Falk's Dual Problem:

$$\begin{array}{ll} \text{Max} & d(\mu) = \text{Min} L(x, \mu) \\ \mu & x \end{array} \quad (1.23)$$

$$\text{s.t.} \quad \mu \geq 0 \quad (1.24)$$

It is easily seen that Wolfe's dual is related to the Kuhn-Tucker conditions while Falk's dual is related to the saddle point conditions. Both dual formulations are useful in developing algorithms as they enable one to solve a primal problem by means of solving an associated dual problem which may be solved more easily than the primal.

1.3 CONSTRAINED OPTIMIZATION METHODS

Over the last three decades, there has been a massive body of published work on the solution of constrained optimization problems. A comprehensive review of all these methods is beyond the scope of this thesis and the reader is referred to Refs.[13], [26], [34], [39], [46], [68] and [71].

At the current stage of research it is not possible to say definitely that one method is entirely preferable or even that the best method has yet been found. Nevertheless, it is increasingly recognized that a good algorithm should be related to the Lagrangean function in one way or the other.

The existing methods of constrained optimization can be roughly put into two categories: (1) primal methods, and (2) transformation methods. There are also many methods in each category. We restrict ourselves in this section only to those methods which are either computationally efficient or theoretically important in the context of the present thesis. Included are two methods, the feasible direction method and the gradient projection method, from the first category and three methods, the penalty function method, the Lagrange method and the multiplier method, from the second category.

A mid-way approach between the primal and transformation methods, which is referred to as the surrogate constraint approach, is examined in Section 1.4 since it is particularly relevant to the present work.

1.3.1 PRIMAL METHODS

By a primal method of solution we mean a search method that works on the original problem directly by searching through the feasible region for the optimum point.

Briefly speaking, the idea is to pick a starting point inside the feasible region and to find a direction, termed a usable feasible direction, such that: (1) a small move in that direction violates no constraint, and (2) the objective function improves in that direction. One then moves some distance in this direction, obtaining a new and better point. The procedure is repeated until a point is obtained such that no usable feasible direction can be found.

An iterative procedure for this method is shown geometrically in Fig.1.4. The search starts with the point x^0 and the usable feasible direction $s^0 = -\nabla f(x^0)$. If a move along s^0 is made to minimize $f(x)$ whose values are depicted by its contours, then the new point is x^1 . Proceeding in the negative gradient direction at x^1 violates the constraints. There are, however, many directions in which one could move at this point, i.e., any direction pointing into the feasible region or along a constraint boundary. Among them, the local "best" direction s is usually chosen which is that feasible direction along which $f(x)$ decreases most rapidly, i.e., along which $-s^T \nabla f(x)$ is maximized. Geometrically, this is the feasible direction making the smallest angle with $-\nabla f(x)$ and is the projection of $-\nabla f(x)$ onto the

constraint boundary. At the point x^1 , this is the vector s^1 . The farthest point one can move along s^1 is x^2 . Repeating the procedure leads to x^3 with $-\nabla f(x^3)$. At this point there is no usable feasible direction in which case x^3 is a minimum point of $f(x)$ over the feasible region.

It should be noted that the above procedure is devised from considering the objective function as a hypersurface and using geometric arguments to devise search strategies. The feasible region is a topological domain defined deterministically by (constraint) function hypersurfaces. These geometric deterministic hypersurfaces play a large part in the development of the procedure. Calculated "information" (function values, gradients, etc.) is used in a geometrical way to search a deterministic topological domain for the optimum point.

In what follows, two primal methods are examined of which one is Zoutendijk's feasible direction method and the other is Rosen's gradient projection method.

Zoutendijk's Feasible Direction Method[107]

The idea of feasible direction methods is to take steps, through the feasible region, of the form

$$x^{r+1} = x^r + \alpha^r d^r \quad (1.25)$$

where d^r is a direction vector, and α^r a nonnegative scalar, being chosen to minimize the objective function $f(x)$ along the direction d^r with the restriction that the point x^{r+1} be feasible. Thus a feasible direction method can be considered as a natural extension of unconstrained descent methods. Each step is the composition of selecting a feasible direction, and a constrained line search. For a feasible point x^r , there can be many feasible directions from which a particular feasible direction is chosen in such a way that the objective function improves as greatly as possible while keeping the constraints satisfied. Zoutendijk proposed determining the feasible direction vector d^r by means of a linear programming subproblem:

$$\text{Min} \quad \nabla f(x^r) d^r \quad (1.26)$$

$$\text{s.t.} \quad a_j^T d^r \leq 0 \quad j \in J_a \quad (1.27)$$

where J_a is the set of indices representing active constraints at x^r and a_j is the coefficient vector of the j th active constraint (or its linearized form). The linear program produces the locally best feasible direction in a sense similar to the steepest descent methods.

Rosen's Gradient Projection Method[82,83]

The gradient projection method is motivated by a desire to implement the feasible direction philosophy while not requiring the solution of a linear program at each step. Instead of attempting to find the direction that leads to the best local improvement while maintaining feasibility, one seeks a direction that is perhaps not quite as good but is much easier to calculate.

Rosen's gradient projection method calculates, at a feasible point x^r , the feasible direction by

$$d^r = -P^r \nabla f(x^r) \quad (1.28)$$

where

$$P^r = I - A^T (AA^T)^{-1} A \quad (1.29)$$

where P^r represents the projection matrix, and A is the Jacobian matrix of the active constraints at x^r . The method requires that the direction vector d^r lies in the tangent subspace defined by the active constraints. The particular direction vector that is used is the projection of the negative gradient $-\nabla f(x^r)$ onto the subspace.

An attractive feature of the primal methods is that they always generate a sequence of feasible points corresponding to decreasing objective functions. Even when the optimization process terminates

before convergence has been achieved, an improved design is generally obtained.

Their main drawback arises from the special treatment of the constraints they require, especially for problems with nonlinear constraints. Keeping them satisfied is an arduous task that always demands a sophisticated algorithm. Consequently, primal methods are effectively used mainly for linearly constrained problems.

In an attempt to alleviate the difficulties of the primal methods, various kinds of transformation methods that convert the original problem into a sequence of unconstrained problems have been proposed. Several transformation methods, such as penalty function methods, Lagrange methods and multiplier methods, are discussed below.

1.3.2 PENALTY FUNCTION METHODS

Penalty function methods[34] transform the original problem into a sequence of unconstrained problems by adding to the objective function a penalty term reflecting the degree of violation of the constraints. In general the penalty functions take the following form:

$$\phi(x, c^r) = f(x) + c^r p(x) \quad (1.30)$$

where $c^r > 0$ is a controlling parameter and $p(x)$ represents the penalty term associated with the constraints, which vanishes if and only if x is feasible. Many different forms of $p(x)$ have been suggested. For instance, where the original problem has only equality constraints, the most commonly used $p(x)$ is:

$$p(x) = \sum_{k=1}^l (h_k(x))^2 \quad (1.31)$$

The procedure for solving the original problem is as follows: let c^r ($r=1,2,\dots$) be an increasing sequence tending to infinity, and solve a sequence of unconstrained problems:

$$\text{Min } \phi(x, c^r) \quad r=1,2,\dots \quad (1.32)$$

for given parameters c^r ($r=1,2,\dots$), obtaining a sequence of minimizing points x^r . It has been proved[34] that under rather general conditions the sequence of minimizing points x^r converges to the solution of the original problem provided c^r is large enough.

For problems with inequality constraints only, a class of penalty functions, termed barrier functions, is usually used. A typical one of these functions is:

$$p(x) = \sum_{j=1}^m (-1/g_j(x)) \quad (1.33)$$

which prevents the solution points x^r from approaching the boundary of the feasible region. These effects will diminish as c^r decreases. The algorithms based on this approach must start with a feasible initial point. An attractive feature of this approach is the fact that the successive unconstrained problems always have their minima inside the feasible region.

When there exist both equality and inequality constraints, mixed penalty functions may be employed which combine the penalty terms of (1.31) with the barrier terms as shown in (1.33).

The penalty function methods have been widely accepted in practice due to the simplicity of the approach, their ability to handle nonlinear constraints, and the availability of very powerful unconstrained minimization methods for solving the problem (1.32). There are, however, considerable disadvantages such as ill-conditioning and slow convergence which arise from the requirement that the controlling parameter c^r must become very large (or very small, as appropriate) to force the minima of the successive unconstrained problems to converge to the minimum of the original problem.

1.3.3 LAGRANGE METHODS

As we have noted earlier, it is impractical to solve a constrained optimization problem by means of directly solving the system of equations and inequalities of the stationarity conditions (1.7) and (1.8) or the Kuhn-Tucker necessary conditions (1.10)-(1.13). A general approach to using Lagrangean functions such as (1.9) is based on sequential minimizations of the Lagrangean function; that is, one minimizes $L(x, \mu^r)$ over x -space for a sequence of multiplier vectors μ^r . This sequence is generated by iterations of the form

$$\mu_j^{r+1} = \mu_j^r + \alpha^r g_j(x^r) \quad j=1, \dots, m \quad (1.34)$$

where x^r is a minimizing point of $L(x, \mu^r)$ over x -space, and α^r is a stepsize parameter.

The iteration (1.34) may be viewed as a steepest ascent iteration aimed at finding an optimum solution of an associated dual problem. Thus this approach is referred to as a primal-dual method which means that alternate iterations over x -space (primal) and μ -space (dual) are required.

In order for the dual problem to be well defined and the iteration (1.34) to be meaningful, a convex structure at the solution point is required. It is usually necessary to minimize the Lagrangean function a large number of times since the ascent iteration (1.34) converges only

moderately quickly. Thus this method has found applications only in the limited class of problems where minimization of the Lagrangean function can be carried out very efficiently due to some special structure of the problem, such as a separable structure.

1.3.4 MULTIPLIER METHODS

Starting around 1968, a number of researchers have proposed a class of methods, called multiplier methods, in which the penalty idea is merged with the primal-dual and Lagrangean philosophies. In the original form, proposed by Hestenes[59] and Powell[77], the methods deal only with equality constrained problems and the quadratic penalty term is added not to the objective function $f(x)$ of the problem but rather to the Lagrangean function, thus forming the augmented Lagrangean function

$$L_r(x, \mu) = f(x) + \sum_{k=1}^l \mu_k h_k(x) + (1/2)c^r \sum_{k=1}^l (h_k(x))^2 \quad (1.35)$$

A sequence of unconstrained minimizations of the form

$$\text{Min } L_r(x, \mu^r) \quad (1.36)$$

is performed where c^r is a sequence of positive penalty parameters. The multiplier sequence μ^r is generated by the iteration

$$\mu_k^{r+1} = \mu_k^r + c^r h_k(x^r) \quad (1.37)$$

where x^r is a solution of the problem (1.36). The initial μ^0 is selected a priori, and the sequence c^r may be either preselected or generated during the computation according to some appropriate scheme.

It turns out that, by combining features of the penalty and the primal-dual approaches, the multiplier methods actually moderate the disadvantages of both. Convergence in the multiplier methods can usually be attained without the need to increase c^r to infinity thereby alleviating the ill-conditioning problem that plagues the penalty method. In addition, the multiplier iteration (1.37) tends to converge to the optimum Lagrange multipliers much faster than the iteration (1.34) of the primal-dual method. Because of these attractive features, the multiplier methods and their subsequently developed variations have emerged as a very important class of constrained optimization methods.

Rockafellar[79,80] has suggested a suitable modification of the above multiplier function to deal with the inequality constrained problem (P). The theoretical and practical details associated with this development closely parallel those for the problem (E). A comprehensive description of the multiplier methods is given by Bertsekas[13].

Nevertheless, the Powell-Hestenes-Rockafellar method requires a sequence of unconstrained minimization problems to be solved and it is the case that more efficient methods can be obtained by avoiding this sequence, for example using exact penalty function. In the present

state of the art, unfortunately, these methods are neither as reliable nor as easy to program.

1.4 SURROGATE APPROACH TO CONSTRAINED OPTIMIZATION

A surrogate problem is one in which the original constraints are replaced by only one constraint, termed the surrogate constraint, which is a positive linear combination of the original constraints. This is, therefore, a midway approach between the primal problem and its unconstrained transformations.

The concept of a surrogate constraint was first introduced by Glover[48] who employed it in solving integer programming problems. Grrenberg and Pierskalla[52] provided the first major theoretical treatment of surrogate constraints in the context of general mathematical programming. There are also other researchers such as Everett[31], Geoffrion[44] and Luenberger[70] who have made important contributions to the literature on surrogate constraints.

For the inequality constrained problem (P) stated by (1.1) and (1.2), referred to as the primal problem, the corresponding surrogate problem has the form:

$$(S) \quad \text{Min} \quad f(x) \quad (1.38)$$

$$\text{s.t.} \quad \sum_{j=1}^m \lambda_j g_j(x) \leq 0 \quad (1.39)$$

where $\lambda_j (j=1, \dots, m)$ are nonnegative weights, termed surrogate multipliers, which may be normalized without loss of generality by requiring

$$\sum_{j=1}^m \lambda_j = 1 \quad (1.40)$$

Before we discuss solution methods for problem (S), it is worthwhile examining further some relationships between problems (P) and (S) in some detail.

1.4.1 SOME RELATIONSHIPS BETWEEN PROBLEMS (P) AND (S)

Some relationships between problems (P) and (S) are presented below as a proposition. Their proofs are straightforward and can be found in the aforementioned papers.

Proposition 1.3:

1. The feasible region of problem (P) is always included by the feasible region of problem (S).
2. If x^S solves problem (S) and x^* solves problem (P), then $f(x^S) \leq f(x^*)$ for all $\lambda \geq 0$.
3. If x^S solves problem (S) and is also feasible in problem (P), then x^S also solves problem (P).
4. If at least one of original constraints is active at the optimum, then the surrogate constraint must be active at the optimum.

Proposition 1.3(1) can be easily verified. If x' is a feasible point in problem (P), i.e., $g_j(x') \leq 0$ ($j=1, \dots, m$), x' also satisfies (1.39) for all $\lambda \geq 0$. Proposition 1.3(1) then follows. A comparison of the primal feasible region with the surrogate one is illustrated in Fig.1.5 which shows that the primal feasible region is always a part of the surrogate feasible regions for any values of λ . Problem (S) can, therefore, be viewed as a relaxation of problem (P). Consequently Proposition 1.3(2) follows. This implies that a (weak) duality relationship exists between problems (P) and (S). From 1.3(1) and 1.3(2), Proposition 1.3(3) holds. This may be used as a criterion for terminating a surrogate algorithm. If the surrogate constraint is inactive at the optimum point, this means that all the original constraints have no effects on the solution. This produces a contradiction of the requirement that at least one of them is active.

Proposition 1.3(4) should hold, and thus the surrogate constraint (1.39) is often written as an equality later on. These relationships are very useful in developing surrogate algorithms.

1.4.2 SURROGATE SOLUTION PROCEDURE

In a surrogate approach to constrained optimization, the desired solution x^* of problem (P) will be sought indirectly through a sequence of solutions of problem (S). Each of the surrogate problems corresponds to a specific vector λ of surrogate multipliers. This approach assumes, therefore, that problem (P) and (S) are equivalent at the solution point; specifically that a set of multipliers λ^* exists and can be found such that x^* which solves problem (S) with λ^* also solves problem (P). Previous researchers have studied this assumption, establishing conditions on its validity. The proof is not repeated here but is similar to existence and optimality proofs for Lagrange multipliers.

An essential condition to be satisfied is that $(x^*, \lambda^*, \alpha^*)$ should be a saddle point of the surrogate Lagrangean L_S which has the form

$$L_S(x, \lambda, \alpha) = f(x) + \alpha \sum_{j=1}^m \lambda_j g_j(x) \quad (1.41)$$

where α is the Lagrange multiplier associated with the surrogate constraint (1.39). The saddle point $(x^*, \lambda^*, \alpha^*)$ must be such that

$$L_S(x, \lambda^*, \alpha^*) \geq L_S(x^*, \lambda^*, \alpha^*) \geq L_S(x^*, \lambda, \alpha) \quad (1.42)$$

In which the left-hand inequality implies minimization over x -space for specified λ and α while the right-hand inequality implies maximization over λ and α space for specified x .

The surrogate approach in general requires alternate iterations in x -space and λ -space. A typical scheme is as follows. An initial set of multipliers λ^0 is chosen and problem (S) is solved to yield corresponding values of x^0 (by minimization, corresponding to the left-hand inequality in (1.42)). The multipliers are then updated to λ^1 (by a maximization process for fixed x^0 , corresponding to the right-hand inequality in (1.42)) and problem (S) is solved again to give x^1 . The process is repeated until the sequence $(\lambda^0, x^0), (\lambda^1, x^1), \dots, (\lambda^r, x^r), \dots$ converges upon a solution of problem (S), and hence also of problem, at (λ^*, x^*) . The major snag in this iterative scheme lies in updating the multipliers λ . This presents many difficulties and is discussed in Chapter 4.

1.4.3 SURROGATE DUALITY

The solution procedure presented above is in fact a primal-dual approach. The saddle point conditions (1.42) assure that this duality relationship, termed surrogate (constraint) duality exists.

A surrogate dual objective can be defined by:

$$s(\lambda) = \{ \underset{x}{\text{Min}} f(x) : \sum_{j=1}^m \lambda_j g_j(x) \leq 0 \} \quad (1.43)$$

which is the minimum value of $f(x)$ at the solution point of problem (S) for fixed $\lambda \geq 0$ which satisfies, of course, the normality condition (1.40).

It has been noted from Proposition (1.3) that problem (S) is a relaxation of problem (P) and $s(\lambda)$ cannot exceed the optimum value $f(x^*)$ of problem (P). With the surrogate constraint (1.39) becoming a more faithful representation of the original constraints, $s(\lambda)$ approaches $f(x^*)$ more closely. Choices of the set of λ that improve the proximity of (S) to (P) - i.e., that provide the greatest value of $s(\lambda)$ - yield the strongest surrogate constraint in a common sense, and motivate the definition of the surrogate dual.

(SD)

$$\text{Max} \quad s(\lambda) \quad (1.44)$$

$$\text{s.t.} \quad \sum_{j=1}^m \lambda_j = 1 \quad (1.45)$$

$$\lambda_j \geq 0 \quad j=1, \dots, m \quad (1.46)$$

Greenberg and Pieskalla[52] have shown that the dual function $s(\lambda)$ is a quasiconcave function, thus assuring that any local maximum for $s(\lambda)$ is also a global maximum for $s(\lambda)$ (disregarding the possibility of a sequence of "plateaus"). In addition, their paper was the first to demonstrate rigorously a smaller duality gap for the surrogate approach than for the Lagrangean approach. It also provided sufficient conditions for the nonoccurrence of surrogate duality gaps. Glover[50] established necessary and sufficient conditions for optimality which invite direct comparison with those for Lagrangean duality. Computationally, an implementation of surrogate duality in posynomial programming problems was made by Dinkel and Kochenberger[24] in terms of geometric programming duality. For more knowledge of the surrogate duality theory, readers should refer to the aforementioned papers.

It should be noted that only in a few cases, for example when the problem functions are separable, the dual function $s(\lambda)$ has an explicit form. For more general cases, it can only be obtained numerically by solving the surrogate problem (S) for fixed λ . The solution of problem (SD) requires, therefore, a sequence of surrogate problems (S) to be solved. This demands an effective scheme to update λ after each surrogate problem (S) is solved. It is obvious that a steepest ascent

iteration similar to Eq.(1.34) is not efficient due to its slow convergence. A better update scheme is therefore needed.

1.5 MOTIVATIONS AND SPECIFICATIONS OF THE PRESENT RESEARCH

In the previous sections, some important theoretical results and computational methods of constrained optimization have been introduced. As stated earlier, difficulties in the solution of constrained optimization problems arise from the presence of many nonlinear constraints. The primal methods are used effectively mainly for linearly constrained problems. The transformation methods make use of powerful unconstrained minimization algorithms and have wider applicability. In particular the multiplier methods are widely accepted since they combine advantages and overcome disadvantages of both conventional penalty methods and Lagrange methods. The main reason for this achievement arises from efficient updates of the Lagrange multipliers such that fast rates of convergence are obtained. However they require sequential unconstrained minimizations and some penalty parameters are updated rather arbitrarily whereas exact penalty functions suffer from instability due to discontinuity of their derivatives. Although there is still debate about which approach is best, it has been widely recognized that a better optimization method should be related in one

way or the other to the Lagrangean function; that is, interactivity among the Lagrange multipliers should be fully taken into account.

The surrogate approach has attracted little attention in nonlinear programming as it requires a sequence of surrogate problems with a single constraint to be solved. However it provides a framework to investigate the interactivity between constraints, i.e., between Lagrange multipliers. The success of the multiplier methods suggests that the interactivity of Lagrange multipliers should be further investigated from both computational and theoretical viewpoints.

1.5.1 MOTIVATIONS OF THE PRESENT WORK

As noted earlier, the geometric and deterministic viewpoint has played a large part in the development of optimization methods. These methods have been devised from considering the objective function as a hypersurface and using geometric arguments to devise a search strategy as illustrated in Fig.1.4. In this geometric interpretation, the constraints are viewed as deterministic boundaries which must not be crossed. Terms such as "gradients", "steepest descent" and "barriers" all have topological associations. They use calculated "information" (function values, gradients, etc.) in a geometrical way to search a deterministic topological domain for an optimum point. Information

theory appears to be incompatible with this as it is essentially concerned with probabilities.

In contrast with conventional methods, a different viewpoint is adopted in the present work. One of its objectives is to set optimization in a non-deterministic context without topological analogies and to use calculated "information" in a different, non-geometrical, information-theoretic way to locate the optimum point.

The idea behind this approach is based upon the speculation that an optimization problem could be thought of as a communication system in which "messages" are received and transmitted alternatively. Design of the system requires that the messages be correctly translated and effectively used subsequently. Thus, methods in information theory can then be used in the solution of optimization problems. It is the hope that this approach would yield new insights into optimizing processes and develop new and radically different algorithms.

From a statistical point of view, a communication system has mathematical similarities to a statistical thermodynamic system. An optimization problem is consequently simulated in the present work as a statistical thermodynamic system for convenience of the later presentations.

1.5.2 SPECIFICATIONS OF THE PRESENT WORK

The present work is the first step towards providing optimization with an information-theoretic basis. The means is to simulate an optimization problem as a statistical thermodynamic system. Several questions are then raised about how to do this simulation. They are:

1. What are micro-states of this statistical thermodynamic system in an optimization context?
2. What are the probabilities of the micro-states?
3. What common characteristic is there in these two processes?
4. What common law governs them?

These questions are briefly answered as follows:

1. Each micro-state in a multi-constrained optimization problem corresponds to a sub-optimization problem in which the objective function is optimized subject in turn to each of the original constraints.
2. The surrogate multiplier associated with each constraint is interpreted as the probability of the system being in the corresponding micro-state.
3. An optimizing process can be thought as a sequence of transitions of the system to its equilibrium states such that the "equilibria" become the common characteristic in the two processes.

4. That the entropy of the system attains a maximum value at an equilibrium state represents the common law to govern the two processes.

To study these objectives, the thesis is arranged in the following way.

1. Two important results in information theory, the Shannon entropy measure and Jaynes' maximum entropy criterion, and some of their applications are examined. Emphasis is placed on those applications which have been made in an optimization context.
2. The entropy maximization process is further investigated to discover its wider applicability and its meaning in an optimization context.
3. The statistical thermodynamic model of optimization is formed such that the methods of information theory can be applied to the solution of optimization problems.
4. The nature of this simulation is further explored and some links between the present approach and existing optimization methods are sought.
5. Some example problems are provided. They are solved by the present approach and the results are compared with those obtained by other methods.
6. The wider applicability of the present approach is investigated.

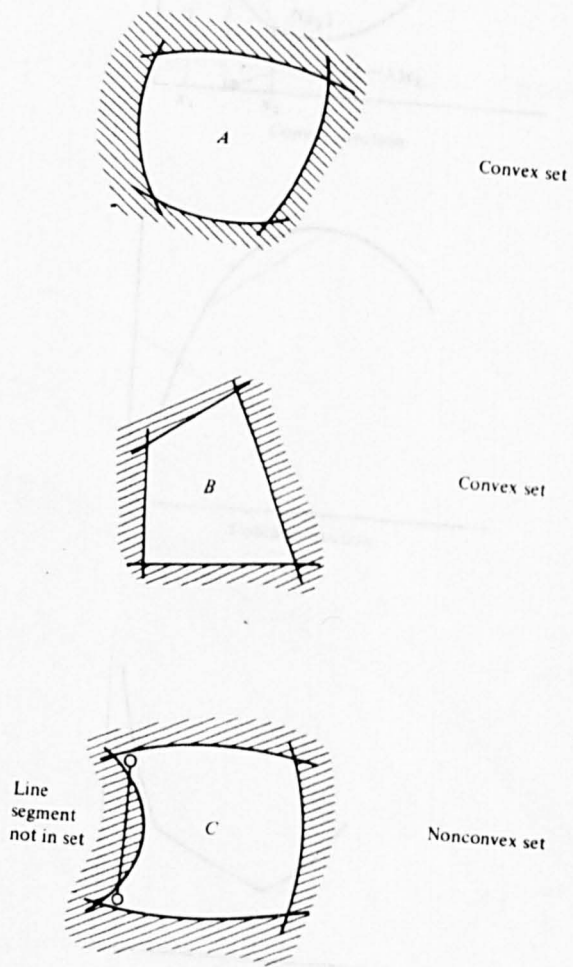


Fig.1.1 Convexity of Sets

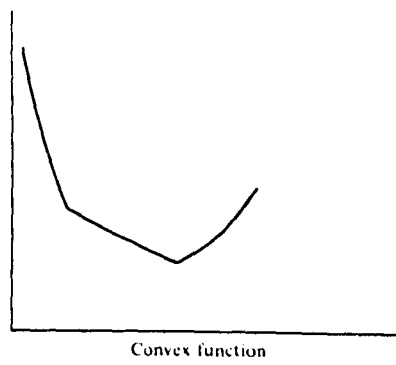
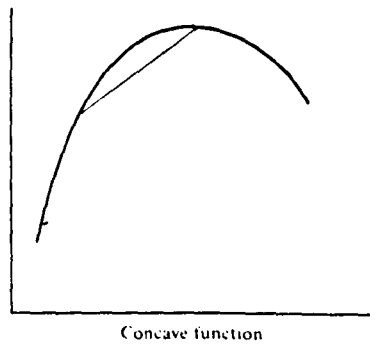
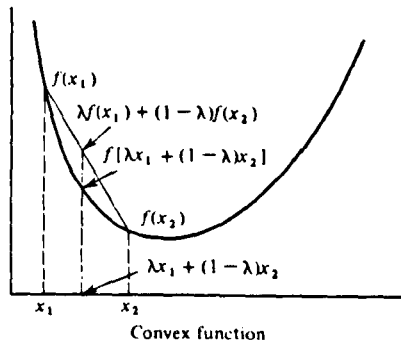


Fig.1.2 Convexity of Functions

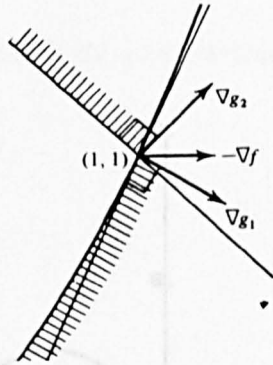


Fig.1.3 Geometry of Kuhn-Tucker Conditions

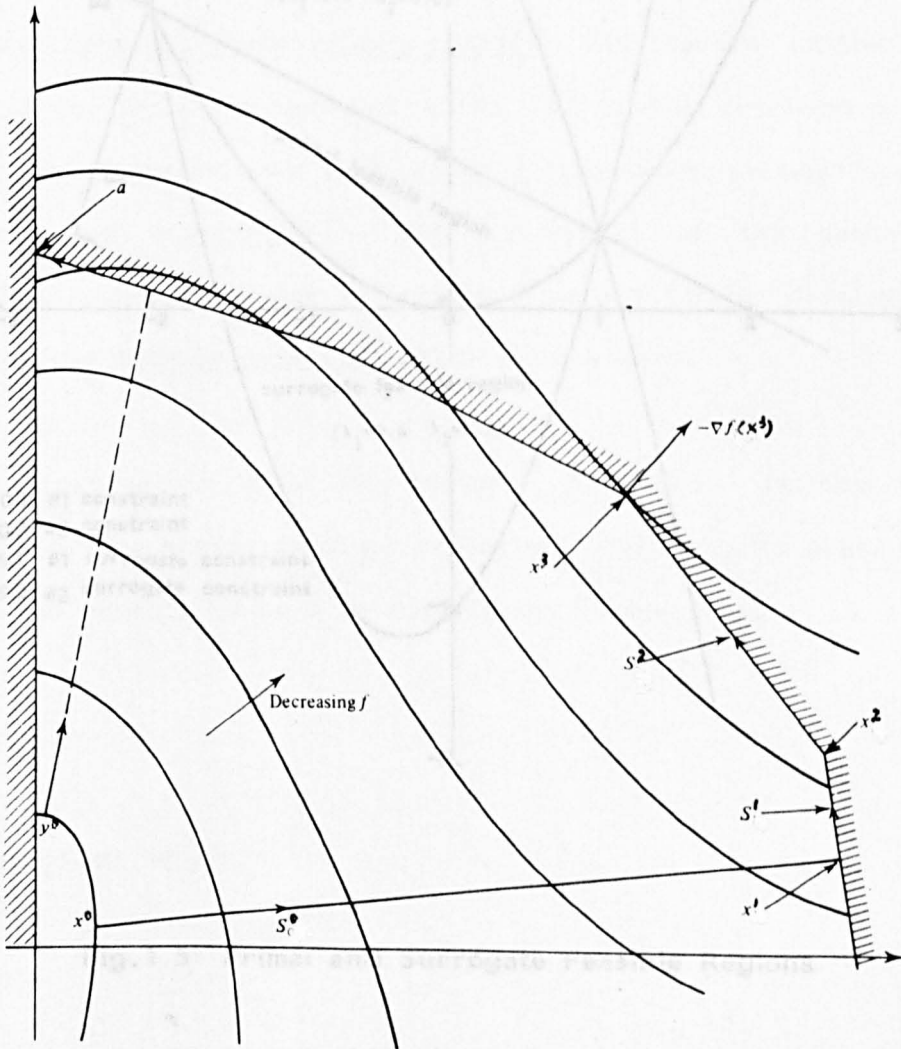
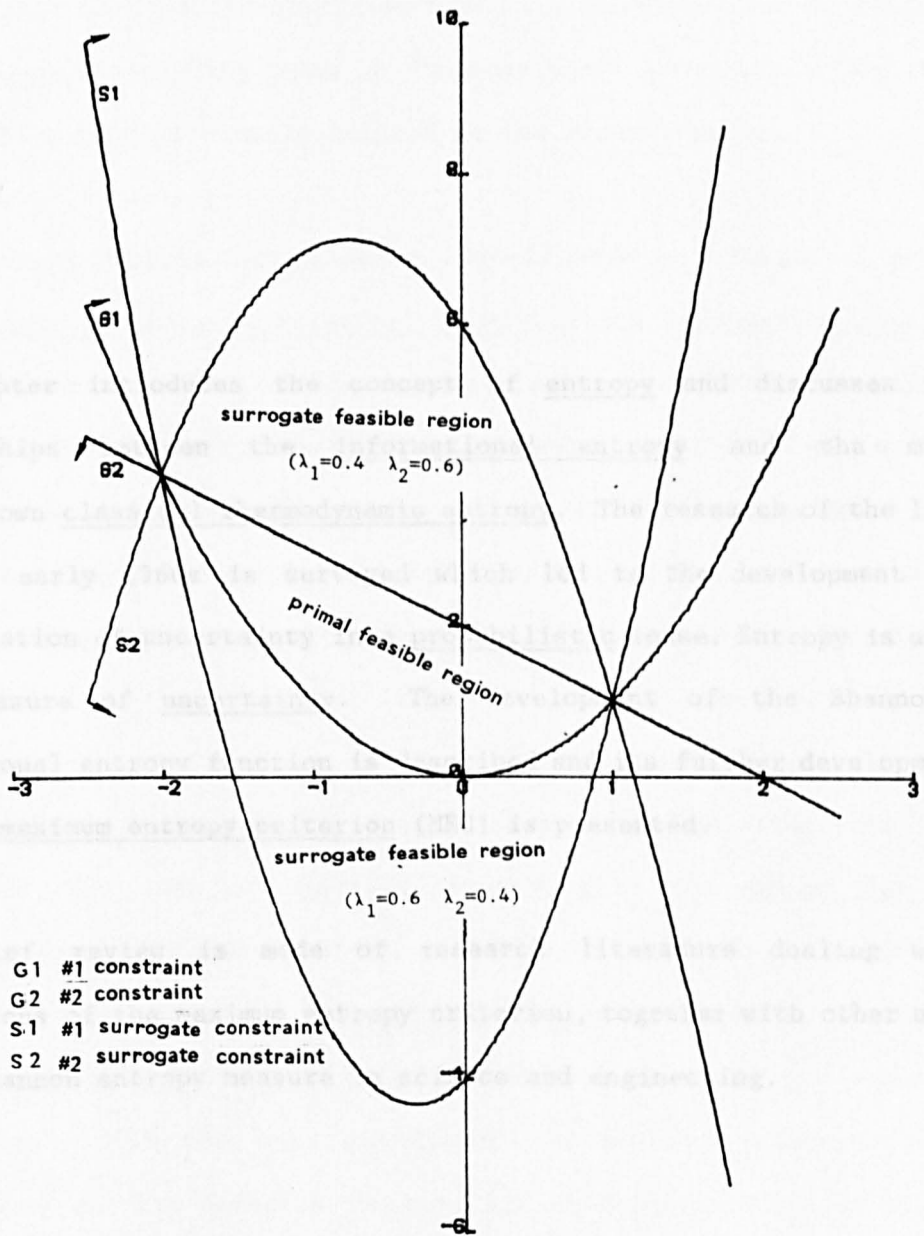


Fig.1.4 Feasible Direction Method

SUMMARY

This chapter introduces the concept of entropy and discusses the relationship between entropy and the maximum entropy criterion. The development of the maximum entropy criterion is traced back to the work of Jaynes in the 1940s to early 1950s. The maximum entropy criterion is used as a measure of uncertainty in a system. The maximum entropy criterion is applied to a wide range of problems in physics, engineering, and the social sciences.



- G1 #1 constraint
- G2 #2 constraint
- S1 #1 surrogate constraint
- S2 #2 surrogate constraint

2.3 INTRODUCTION

Fig.1.5 Primal and Surrogate Feasible Regions

As stated in the introductory chapter on optimization, the maximum entropy criterion is used as a measure of uncertainty in a system. The maximum entropy criterion is applied to a wide range of problems in physics, engineering, and the social sciences.

CHAPTER 2 ENTROPY AND THE MAXIMUM ENTROPY CRITERION

SUMMARY

This chapter introduces the concept of entropy and discusses the relationships between the informational entropy and the much better-known classical thermodynamic entropy. The research of the late 1940s to early 1960s is surveyed which led to the development and quantification of uncertainty in a probabilistic sense. Entropy is used as a measure of uncertainty. The development of the Shannon's informational entropy function is described and its further development into the maximum entropy criterion (MEC) is presented.

A brief review is made of research literature dealing with applications of the maximum entropy criterion, together with other uses of the Shannon entropy measure in science and engineering.

2.1 INTRODUCTION

As stated in the introductory chapter, an optimization problem is simulated in the present work as a statistical thermodynamic system

while an optimizing process is thought of as transitions of the system to a sequence of equilibrium states which are characterized by certain entropy maxima depending upon a "temperature" parameter. Thus the concept of entropy is closely related to the present study.

The entropy concept has played a central role in a number of areas such as thermodynamics, statistical mechanics and information theory. It springs from two roots. On the one hand, in classical thermodynamics, entropy is defined as a macroscopic thermodynamic variable of the system under consideration; on the other hand, in statistical mechanics and information theory, it is defined as a measure of the number of ways in which components of a system may be arranged under given circumstances.

In classical thermodynamics, this concept has a significance no less fundamental than that of energy. According to the second law of thermodynamics, the entropy in an isolated system tends to a maximum so that this variable is a criterion for the direction in which processes can take place. On the relation between energy and entropy, Emden wrote: 'In the huge manufactory of natural processes, the principle of entropy occupies the position of manager, for it dictates the manner and method of the whole business, whilst the principle of energy merely does the book-keeping, balancing credits and debits' (See Fast[33]).

Classical thermodynamics in which the entropy concept originated is concerned only with the macroscopic states of matter, i.e. with the

experimentally observable properties. It thus does not enquire into the mechanisms of phenomena and is therefore unconcerned with what happens on a microscopic scale. The microscopic picture, however, can help to give deeper meanings to the thermodynamic laws and concepts. The branch of science concerned with this aspect is statistical mechanics.

When a small number of macroscopic variables such as pressure, temperature, volume, chemical composition, etc. of a system are known, the thermodynamic state of this system is known. It is clear that a description of this kind still leaves open many possibilities as regards the detailed state on an microscopic scale. One and the same state, in the thermodynamic sense, thus comprises very many states on the microscopic scale; that is, a thermodynamic state can be realized in many ways or micro-states. If the number of these micro-states is denoted by m , then entropy of the system is defined as

$$S = k \ln m \quad (2.1)$$

where k is Boltzmann's constant. The quantity S in Eq.(2.1) may be considered as the statistical interpretation of entropy.

That entropy tends to a maximum means, according to (2.1), a tendency towards the state with a maximum number of possibilities of realization, i.e. a tendency towards the most probable state. In order to be able to apply (2.1) directly, all the micro-states must have the same probability of occurring. A definition of the entropy with more general validity than Eq.(2.1) will be introduced in Section 2.2.

Entropy of a system was first defined by Clausius as a function of some macroscopic variables that can be directly measured. The Clausius' entropy is a nonprobabilistic concept and is usually referred to as the classical entropy. Boltzmann was the first to emphasize the probabilistic meaning of the entropy. He noticed that the entropy of a physical system can be considered as a measure of "disorder" in the system and that in a system having many degrees of freedom, the number measuring the disorder of the system measures also the uncertainty about individual micro-states. However, he made no explicit reference to "information" and his entropy is, therefore, referred to as the statistical entropy. It was Shannon[86] who first introduced the entropy concept as a measure of uncertainty or information in an information theory context. Thus Shannon's entropy is referred to as the informational entropy which has wider applicability than the statistical entropy.

2.2 DEFINITIONS AND PROPERTIES OF INFORMATIONAL ENTROPY

One of the fundamental building blocks of modern information theory is the paper by Shannon[86] in which a new mathematical model of communication systems was proposed and investigated. The most important innovation of this model was that it considered the components of a communication system as probabilistic entities. In his paper, Shannon

proposed a quantitative measure of the amount of uncertainty about the possible outcomes of a probabilistic experiment.

Consider a probabilistic experiment having n discrete possible outcomes a_1, \dots, a_n with the respective discrete probabilities p_1, \dots, p_n , satisfying the following axiomatic conditions

$$p_i \geq 0 \quad (i=1,2,\dots,n) \quad \text{and} \quad \sum_{i=1}^n p_i = 1$$

Such an experiment, of course, contains an amount of uncertainty about the particular outcome which will occur if we perform the experiment. It can be seen that this amount of uncertainty, contained a priori by the probabilistic experiment, essentially depends on the probabilities of the possible outcomes of the experiment. For instance, if we have a probabilistic experiment having only two possible outcomes a_1, a_2 with two different sets of probability distributions $(p_1=0.5, p_2=0.5)$ and $(p_1=0.96, p_2=0.04)$, it is obvious that the first case contains more uncertainty than the second. In the second case, the result of the corresponding experiment is "almost surely" a_1 , while in the first case we cannot make any prediction on the particular outcome which will occur. This shows that a uniform probability distribution has a larger amount of uncertainty associated with it than a non-uniform distribution.

Shannon was able to postulate a measure of such uncertainty on a quantitative basis in the following way. He proposed that a measure for uncertainty should satisfy the following requirements:

1. It should be continuous in the p_i ($i=1, \dots, n$).
2. If all the p_i are equal, then it should be a monotonically increasing function of the number of outcomes n . This reflects the need that with equally possible outcomes, the measure of uncertainty should be higher when there are more possible outcomes than when there are few.
3. The uncertainty about two independent events A and B should be the sum of the uncertainties about A and B taken separately.

Shannon demonstrated that these criteria were sufficient to define uniquely the function:

$$S = -k \sum_{i=1}^n p_i \log p_i \quad (2.2)$$

where k is merely a positive constant depending on a suitable choice for the units of measure, and it is defined that $0 \log 0 = 0$. The function S in Eq.(2.2) is referred to as the informational entropy which is the form of entropy used in this thesis.

Shannon's entropy measure has a number of properties which might be expected of a reasonable measure of uncertainty in a probabilistic context. They are summarized as the following proposition:

Proposition 2.1:

1. $S_n(p_1, \dots, p_n) \geq 0$
2. If $p_k=1$ and $p_i=0$ ($i=1, \dots, n, i \neq k$), then $S_n(p_1, \dots, p_n)=0$
3. $S_{n+1}(p_1, \dots, p_n, 0) = S_n(p_1, \dots, p_n)$
4. $S_n(p_1, \dots, p_n) \leq S_n(1/n, \dots, 1/n)$
5. Shannon's entropy function S in Eq.(2.2) is a continuous and symmetric concave function w.r.t. all its arguments.

The proof of Proposition 2.1 is very straightforward from Eq.(2.2) and can be found elsewhere[54]. According to Proposition 2.1(2), the entropy S is equal to zero if one of the probabilities p_1, \dots, p_n is unity and all others are zero. But this is just the case where the result of the experiment can be predicted beforehand with complete certainty, so that there is no uncertainty about the outcome. From Proposition 2.1(4), the probabilistic experiment with the greatest uncertainty is the one with equally likely outcomes; that is, Shannon's entropy measure, for fixed n , assumes its greatest value for the uniform p.d. The concavity of the Shannon entropy function S , according to Proposition 2.1(5), is a nice property which enables Jaynes' maximum entropy problem to be a convex programming problem as will be discussed in the next section.

2.3 THE MAXIMUM ENTROPY CRITERION (MEC)

Shannon's entropy measure was an important step forward in that it allowed the amount of uncertainty in a probabilistic experiment to be quantified provided that the probabilities of all outcomes are known. The next important advance was made by Jaynes[60,61,63] who realized that in many probabilistic experiments the probabilities of discrete outcomes are often not known (unknown prior probabilities). Jaynes extended the use of the Shannon entropy measure to calculate the unknown prior probabilities from observable data on the probabilistic experiment, and hence extended the role of Shannon's function (2.2) from a simple measure to a crucial role in an inference process; i.e. given a probabilistic process and observed aggregated data from that process, what does the Shannon measure of uncertainty allow us to logically infer about the p.d's underlying the process?

Suppose there exists an observable probabilistic process in which a discrete random variable can take on any one of n outcomes of value x_1, \dots, x_n . Suppose also that as a result of observations on the process it can be deduced that the outcomes satisfy certain aggregated functional relationships $g_1(x), g_2(x), \dots$, such as mean values, variance, etc. Let there be m such functions where $m \ll n$. What can be deduced about the probabilities p_1, \dots, p_n of the random variable attaining values x_1, \dots, x_n ? Clearly an infinite number of p.d's can satisfy the m observed functions g_1, \dots, g_m . Which one should be chosen? Some selection criterion is then needed.

Jaynes, in a brilliant paper[60], wrote: 'in making inference on the basis of partial information we must use that probability distribution which has maximum entropy subject to whatever is known. This is the only unbiased assignment we can make; to use any other would amount to arbitrary assumption of information which by hypothesis we do not have.'

Jaynes therefore recognized that the infinite number of p.d's which satisfy the m observable behaviour functions all have different Shannon entropy measures of uncertainty. The one with the highest entropy value should be chosen as this introduces minimum artificial bias into the choice.

Jaynes referred to the above statements as the principle of maximum entropy, implying that all else is consequent upon its fundamental nature. In fact, Jaynes established a selection criterion among the p.d's compatible to the given information. The criterion has a subjective character by nature, but it can rather be considered as the most "objectively subjective" criterion for constructing the random distributions.

The Jaynes' selection criterion is, throughout this thesis, referred to as the maximum entropy criterion (MEC) rather than the "principle" as it actually establishes an inference criterion.

Mathematically, to maximize the entropy S of (2.2) subject to the given information, leads to a mathematical optimization problem:

(MEC)

$$\text{Max } S = -k \sum_{i=1}^n p_i \ln p_i \quad (2.2)$$

$$\text{s.t. } \sum_{i=1}^n p_i = 1 \quad (2.3)$$

$$\sum_{i=1}^n p_i g_j(x_i) = E[g_j] \quad j=1, \dots, m \quad (2.4)$$

where $E[.]$ is the expectation operator and it is axiomatic that $p_i \geq 0$. The constant k is set to unity below, which has no effect on the solution.

Problem (MEC) can be solved using the Lagrange multiplier method. Introducing Lagrange multipliers (μ_0+1) for (2.3) and μ_j ($j=1, \dots, m$) for (2.4), we have the Lagrangean function

$$L(p, \mu) = - \sum_{i=1}^n p_i \ln p_i + (\mu_0+1) \left(\sum_{i=1}^n p_i - 1 \right) + \sum_{j=1}^m \mu_j \left(\sum_{i=1}^n p_i g_j(x_i) - E[g_j] \right) \quad (2.5)$$

for which the stationarity conditions w.r.t. p_i are:

$$\partial L / \partial p_i = -(1 + \ln p_i) + (\mu_0+1) + \sum_{j=1}^m \mu_j g_j(x_i) = 0 \quad i=1, \dots, n \quad (2.6)$$

from which

$$\ln p_i = \mu_0 + \sum_{j=1}^m \mu_j g_j(x_i)$$

hence

$$p_i = \exp[\mu_0 + \sum_{j=1}^m \mu_j g_j(x_i)] \quad i=1, \dots, n \quad (2.7)$$

It should be noted that a closed form of the solution to problem (MEC) has been obtained as a result of the separability of the problem functions. The solution contains $m+1$ arbitrary constants, i.e., the $m+1$ undetermined Lagrange multipliers. These multipliers must be chosen in such a way as to satisfy the equality constraints (2.3) and (2.4). When the Lagrange multipliers have been chosen, the entropy function S gives the maximum value consistent with the given data, and therefore it represents the least biased p.d. which is referred to as the maximum entropy distribution (MED) later.

If we substitute p_i from (2.7) into the normality condition (2.3), μ_0 can be expressed in terms of the multipliers μ_j :

$$\exp(-\mu_0) = \sum_{i=1}^n \exp[\sum_{j=1}^m \mu_j g_j(x_i)] \quad (2.8)$$

On substituting (2.8) into (2.7), the probabilities p_i are eventually expressed in terms of the multipliers μ_j only, i.e.,

$$p_i = \exp[\sum_{j=1}^m \mu_j g_j(x_i)] / Q \quad i=1, \dots, n \quad (2.9)$$

where

$$Q = \exp(-\mu_0) = \sum_{i=1}^n \exp\left[\sum_{j=1}^m \mu_j g_j(x_i)\right] \quad (2.10)$$

which is referred in statistical mechanics to as the partition function. The remaining multipliers μ_j are found by solving the m equations (2.4) following the substitution of the solution (2.9). The solution of the system of m equations in μ_j is, however, not easy as these equations are nonlinear and can usually only be solved numerically. This point will be discussed in detail in Chapter 3.

Problem (MEC) and its solution (2.9) are written in terms of a discrete r.v. A similar formulation in the case of continuous r.v.'s now follows. The solution is obtained following the same process as for discrete r.v.'s. The equivalent forms for a continuous r.v. are,

(MECC)

$$\text{Max} \quad S = -k \int f(x) \ln[f(x)] dx \quad (2.11)$$

$$\text{s.t.} \quad \int f(x) dx = 1 \quad (2.12)$$

$$\int g_j(x) f(x) dx = E[g_j] \quad j=1, \dots, m \quad (2.13)$$

as the equivalent of problem (MEC) and

$$f(x) = \exp\left[\sum_{j=1}^m \mu_j g_j(x)\right] / \int \exp\left[\sum_{j=1}^m \mu_j g_j(x)\right] dx \quad (2.14)$$

as the form of its solution, $f(x)$ is the density function of continuous r.v. x , and the integrations are carried out over the range of x .

This thesis is concerned only with the discrete case.

2.4 INTERPRETATION OF THE MAXIMUM ENTROPY CRITERION

Tribus[99] has shown that analytical p.d.'s such as the normal, gamma and exponential distributions are merely different particularizations of the MED. For instance, if the mean value and the variance of some r.v. are given, the MEC gives a normal (Gaussian) distribution. Thus, the normal distribution contains the largest amount of uncertainty compatible with the observed data and is therefore a least biased fit to this data. One valuable aspect of the MEC is that it allows some well known distributions to be interpreted in new lights.

The maximum entropy criterion may be regarded as an extension of Laplace's principle of "insufficient reason" to which it reduces in the case in which no information is given except the enumeration of the possible outcomes x_i . In fact, the MEC is a statement that, in a situation of limited knowledge, we will select the p.d. that is as close to a uniform distribution as our knowledge allows; that is, applying the MEC, we select the most uncertain or "spread out" random distribution subject to the given data. Thus the concept of entropy

supplies a selection criterion and enables the least biased p.d.'s to be found which fit all available information.

2.5 SOME APPLICATIONS OF THE MAXIMUM ENTROPY CRITERION

Shannon's entropy measure and Jaynes' maximum entropy criterion have found applications in a variety of areas of science and engineering. We examine only a few of these areas such as thermodynamics, statistical mechanics, civil engineering, queueing theory, transportation planning, etc. Among them, the transportation problem together with other uses of entropy in an optimization context is left to the next section because of their particular relevance to the present study.

As implied by Jaynes's paper[60], statistical mechanics has been the most fruitful discipline for applications of the MEC. Jaynes stresses that informational entropy is a primitive concept more fundamental even than energy, and he has derived many well-known relations in statistical mechanics in a very elementary way without consideration of "ensembles" or appeal to the usual arguments concerning "ergodicity" or equal a priori probabilities. Thus the principles and mathematical methods of statistical mechanics can be seen to be of much more general applicability than is suggested by conventional arguments.

Another field in which the Shannon's entropy and the Jaynes' MEC have found immediate applications is classical thermodynamics. Following Jaynes, Tribus demonstrated[97,98] that all of the laws of classical thermodynamics, and in particular, the concepts of heat and temperature, could be defined from Shannon's entropy using Jaynes' maximum entropy criterion. Classical thermodynamics is unsatisfactory in that it renders the axiomatic structure more complicated than it needs to be. Attempts to develop clean formulations of thermodynamics, based upon abstract axioms, were made by Caratheodory who succeeded in this objective only partially. It is Shannon's and Jaynes' work that makes it possible to improve the axiomatic structure. The particular formulations were given by Tribus[98] and Callen[17]. In their textbooks, the Shannon entropy measure and Jaynes' MEC are considered as axioms (or postulates) from which the laws of thermodynamics are derived.

Recently, civil engineering has also become an active area in which there have appeared many applications of informational entropy. Basu and Templeman[5], and Siddall and Diab[87] have used the MEC in structural reliability analysis and probabilistic design. In traditional approaches to reliability analysis an analytical p.d. must be chosen by the analyst/designer to represent random loads or strengths. Any bias or lack of fit between the chosen distribution and the available data will be magnified in the extrapolation of the distribution to the tail regions which are important in reliability analysis, as the probability of failure is heavily related to the overlapping tails of load and strength distributions. The MED, as the least biased p.d., is a more appropriate substitute for the analytical

distributions. Basu and Templeman[5] argue that by fitting a MED to available data itself by means of the MEC a more logical and rigorous approach to structural reliability analysis results.

Munro and Jowitt[72] used the MEC in decision analysis in the ready-mixed concrete industry. The principal problem is that of making optimal decisions in the face of the uncertainty associated with the unknown state of future orders. They argued that the evaluation of prior probabilities for the order states should be made objectively and not be affected by any personal bias. Thus in their approach the MEC was used to produce the least biased p.d. associated with the orders for each mix.

In addition Guiasu[55] has also used the MEC in queueing theory. For some single-server queueing systems, when the expected number of customers is given, the MEC gives the same p.d. of the possible states of the system as the birth-and-death process applied to an M/M/1 system in a steady-state condition. For other queueing systems, such as M/G/1 for instance, the MEC gives a simple p.d. of possible states, while no closed-form expression for such a p.d. is known in the general framework of a birth-and-death process. In this work Guiasu argues that use of the MEC strengthens belief that fundamental assumptions in the birth-and-death process which is axiomatic to such queueing systems are correct and justifiable. Furthermore it strengthens some of the initial axioms and improves insight into the problems.

The MEC has also been applied to many diverse areas such as psychology, biology, population forecasting, medicine, search theory, etc. The reader is referred to the lecture papers collected in the book "The Maximum Entropy Formalism" [69].

2.6 APPLICATIONS OF ENTROPY TO OPTIMIZATION PROBLEMS

Many researchers have used Shannon's entropy in the transportation problem, which is generally formulated as an LP problem (Refs.[23], [25], [28-30], [104]). A typical transportation problem is to predict values of a set of variables x_{ij} which represents some spatial interaction between zones i and j (trip distribution). Suppose we are given the totals of flows leaving each zone i , O_i , and entering each zone j , D_j , and the cost of travel between each zone, c_{ij} , then an LP model to minimize the total travelling cost takes the form:

(ST)

$$\text{Min } C = \sum_{ij} c_{ij} x_{ij} \quad (2.15)$$

$$\text{s.t. } \sum_j x_{ij} = O_i \quad (2.16)$$

$$\sum_i x_{ij} = D_j \quad (2.17)$$

and

$$x_{ij} \geq 0 \quad (2.18)$$

which is referred below to as the standard transportation (ST) problem.

Erlander[28,29], among others, adds an entropy constraint

$$-\sum_{ij} x_{ij} \ln(x_{ij}) \geq H \quad (2.19)$$

to the problem (ST) in order to preserve a desired level of "interactivity" (or "accessibility") in the solution to the trip distribution problem where H is a specified value of the entropy of the variables x_{ij} . This modified problem with an entropy constraint is referred to as the entropy-constrained transportation (ECT) problem.

The procedures for solving problems (ST) and (ECT) are not repeated here but can be found in the aforementioned papers. It is interesting that by adding the entropy constraint (2.19), an analytical solution to problem (ECT), expressed in terms of the Lagrange multipliers associated with the constraints (2.16)-(2.19), is obtained. It turns out that the analytical solution is of the well-known gravity model type. Evans[30] has formally proved that the solution to problem (ST) is the limit of the gravity model solution as the Lagrange multiplier associated with the entropy constraint (2.19) approaches infinity.

Erlander[29] has stressed some advantages of this modified model over problem (ST). First of all, the solution of problem (ST) is replaced by the "smoother" solution of problem (ECT). This character may be advantageous in some planning problems where the uncertainty of the real

world has been replaced in the models by deterministic relationships. The entropy constraint may in such cases be viewed as a substitute for lost complexity. The solution obtained has also high "accessibility" since all variables in the solution of problem (ST) that may be zero, become strictly positive in problem (ECT). Thus adding the entropy constraint has made the mathematical modelling more sensible. From a computational point of view, the obtainability of an analytical solution in terms of the Lagrange multipliers has provided an alternative means of solving more general LP problems.

Erlander[29] has extended the entropy constraint approach to solve the following LP problem:

(LP)

$$\text{Min } c^T x \quad (2.20)$$

$$\text{s.t. } Ax = b \quad (2.21)$$

$$\sum_{i=1}^n x_i = 1 \quad \text{and} \quad x \geq 0 \quad (2.22)$$

where c is a given n -vector, x is an n -vector of variables, A is an $m \times n$ coefficient matrix and has full rank ($=m$), and b is an m -vector of resources. By adding an entropy constraint

$$\sum_{i=1}^n x_i \ln x_i \geq H \quad (2.23)$$

to problem (LP), he examines the solution to this modified problem. It is obvious that the solution depends upon values of H. When H is less than a certain value H_{\min} the entropy constraint is slack and therefore has no effect on the solution. For $H > H_{\min}$, the entropy constraint becomes active and the objective function increases as H increases until some maximum value H_{\max} is reached. A further increase of H makes the problem infeasible. Hence, the values of H which are of interest lie in a certain interval $[H_{\min}, H_{\max}]$, where the entropy constraint is active and there are feasible solutions. For values of H in this range, an analytical form of the solutions can be obtained in terms of $m+1$ Lagrange multipliers. Problem (LP) is then reduced to a dual problem in $m+1$ Lagrange multipliers. Erlander proceeds to use Newton-type method to obtain an iterative procedure for the Lagrange multipliers. He has also shown that for proper choices of c, d and H, the entropy constrained problem is equivalent to the following entropy maximization problem:

$$\text{Max} \quad -\sum_{i=1}^n x_i \ln x_i \quad (2.24)$$

$$\text{s.t.} \quad Ax = b \quad (2.21)$$

$$\sum_{i=1}^n x_i = 1 \quad \text{and} \quad x \geq 0 \quad (2.22)$$

and

$$c^T x \leq d \quad (2.25)$$

where $-\sum_{i=1}^n x_i^* \ln x_i^* = H$, $c^T x^* = d$ and $H_{\min} \leq H \leq H_{\max}$. In fact the equivalence can be easily verified if their Lagrangeans are examined.

Another use of entropy in an optimization context is given by Cerny[18]. He has proposed a Monte Carlo algorithm to approximately solve the travelling salesman problem. The algorithm simulates the behaviour of a thermodynamic system that is equivalent to the travelling salesman problem in the following sense. Each permutation of the stations on the travelling salesman trip is considered to be the configuration of the thermodynamic system. The length of the trip then corresponds to the energy of the system in that particular configuration. He argued that the thermodynamic system at a given temperature should spontaneously approach its equilibrium state characterized by a certain mean value of energy, and by simulating the transition of the system to equilibrium and decreasing the system temperature one could find smaller and smaller values of the mean energy (length of the trip). He has proposed a heuristic algorithm based on the above ideas. Several example problems are presented. Cerny's approach is very instructive in that it solves an optimization problem by analogy with thermodynamic equilibria.

Ben-Tal[8] has used entropy in NLP problems with stochastic constraints. The stochastic program is replaced by a deterministic program by penalizing solutions which are not feasible in the "mean". The penalty term is given in terms of a relative entropy functional and is accordingly called an entropic penalty.

However, it should be noted that uses of entropy in the solution of optimization problems are still limited and some of these applications have little rigorous background. The major difficulties in extending

their applications in an optimization context lies in the fact that at present almost all optimization problems are viewed in terms of a geometric, deterministic topological domain. This has overshadowed any other possible interpretations and has so pervaded and preconditioned the thought processes that it is difficult to isolate logical arguments which lead rigorously to an alternative probabilistic and entropic interpretation.

2.7 CONCLUSIONS AND RELEVANCE TO THE PRESENT WORK

In the previous sections, Shannon's entropy measure, Jaynes' maximum entropy criterion (MEC) and their applications to several areas in science and engineering have been examined. In particular, emphasis has been placed on those applications which have been made in an optimization context. Several conclusions can be drawn from this examination.

1. Shannon's informational entropy has wider applicability than both classical entropy and statistical entropy as it is defined in a more general context.
2. Jaynes' maximum entropy criterion is indeed a powerful tool for inference on the basis of partial information and suggests a much more widespread potential for use of the MEC in more application areas than hitherto.

3. The use of entropy in an optimization context and the analogy between an optimizing process and thermodynamic equilibrium are subjectively sensible.

The present study aims to explore the use of Shannon entropy and Jaynes' maximum entropy criterion in an optimization context. The first difficulty to be overcome is to generalize the concept of probability which is invariably required for the use of entropy; that is, appropriate quantities in optimization problems must be sought which could be reasonably interpreted as probabilities.

The optimization problem to be studied in this thesis is the inequality constrained problem (P), as defined by (1.1) and (1.2). It is also assumed that at least one of the original constraints is active at the optimum solution; that is, it is a multi-constrained, non-degenerate NLP problem.

It has been seen in Section 1.4 that if the optimum surrogate multipliers λ^* are known, the solution x^* to problem (P) can be obtained by solving the corresponding surrogate problem (S) once. Unfortunately, it is impossible to predict these optimum multipliers in advance using conventional geometric arguments. The wide applicability of Shannon's entropy and Jaynes' maximum entropy criterion, as demonstrated by the previous applications described in this chapter, suggests that they could help to resolve this difficulty. To this end, the original problem is divided into m , the number of original constraints, subproblems of which each has the form:

(PJ)

$$\text{Min} \quad f(x) \quad (2.26)$$

$$\text{s.t.} \quad g_j(x) \leq 0 \quad (2.27)$$

Each subproblem thus consists of minimizing the original objective function subject in turn to one of the original constraints. Each of these subproblems may be thought of as a micro-state of the whole thermodynamic system, which corresponds to the original problem (P). It is then argued that the most probable state of the system is one in which the entropy of the system achieves a maximum value. This approach is described in detail in Chapter 4.

CHAPTER 3 DUALS OF THE MAXIMUM ENTROPY PROBLEM

SUMMARY

This chapter constructs mathematical dual forms of the maximum entropy problem (MEC) considered in Chapter 2 and develops a computational approach which enables discrete or continuous maximum entropy distributions to be calculated relatively easily. The dual forms give additional insights into the nature of entropic processes. The relationships between informational entropy and experimental thermodynamic entropy are discussed.

Much of the original material of this chapter has been published in a joint paper by the author and Dr.A.B.Templeman[92] entitled "Entropy Duals" which is reproduced as Appendix B of this thesis.

3.1 INTRODUCTION

Shannon's entropy measure and Jaynes' maximum entropy criterion have found throughout science and technology many and widespread applications, as described in Chapter 2, of which uses of entropy in

an optimization context are of particular interest in this thesis. Before examining this main theme, it is worthwhile examining further the nature of entropic processes. This may strengthen our understanding and thereby lead to the discovery of more application areas than hitherto.

It has been pointed out in the last chapter that the Jaynes' maximum entropy problem (MEC), defined by (2.2)-(2.4), is a convex programming problem. This convex structure of problem (MEC) implies the existence of a mathematical dual form. In fact, many researchers(Refs.[1], [97], [98]) have used the dual function in either theoretical or computational aspects. To the author's knowledge, however, they have not formally recognized it as a dual form of entropy. Consequently this has led to some confusions between the two forms of entropy, the primal and the dual.

Two ways of deriving the dual are presented of which one uses the classical Lagrangean Duality theory and the other makes use of the well-established duality theory of geometric programming. Although the two duals appear to be different, they are shown to be equivalent under variable transformation. The dual formulations given herein are in terms of discrete entropy but similar dual forms for continuous entropy are also presented on a pragmatic basis.

An additional aspect of the dual forms is that they afford a simple means of calculating MED's for complex problems using only unconstrained minimization algorithms. A numerical example is presented.

3.2 DUAL FORMS OF THE MAXIMUM ENTROPY PROBLEM (MEC)

As noted earlier, problem (MEC) is a convex programming problem for which there must consequently exist a mathematical dual form. Two ways of deriving the dual are presented below one of which uses Lagrangean duality and the other uses geometric programming duality as described in Appendix A.

3.2.1 LAGRANGEAN DUAL FORMULATION

All the necessary ingredients for deriving a formal dual form of problem (MEC) have been assembled in Section 2.3. A closed form of the solution of problem (MEC), as expressed in Eq.(2.9), has been obtained in terms of m Lagrange multipliers μ_j . Because of the convex structure of the problem, this solution is a global maximum. The solution procedure is not repeated here.

We proceed in deriving the dual by directly substituting μ_0 from Eq.(2.8) and p_j from Eq.(2.9) into the Lagrangean function (2.5). After algebraic manipulation, the dual problem (DE) is obtained as

(DE)

$$\text{Min } D(\mu) = \ln Q - \sum_{j=1}^m \mu_j E[g_j] \quad (3.1)$$

where

$$Q = \sum_{i=1}^n \exp\left[\sum_{j=1}^m \mu_j g_j(x_i)\right] \quad (3.2)$$

Before examining problem (DE) and its relationship to problem (MEC) in detail, another way of deriving a dual problem for problem (MEC) using geometric programming is presented.

3.2.2 GEOMETRIC PROGRAMMING DUAL FORMULATION

Geometric programming was developed in the late 1960s and has become an established technique of mathematical programming. Appendix A to this thesis contains the mathematical forms of GP as used in this thesis. Duffin, Peterson and Zener[26] give a comprehensive background to the topic and full proofs of all the duality relationships which are used here without proof.

To use the primal-dual relationships of GP as stated in Appendix A, a modified form of problem (MEC) is introduced which has a different form of the objective function. From problem (MEC),

$$S = - \sum_{i=1}^n p_i \ln p_i = \sum_{i=1}^n \ln(1/p_i)^{p_i} = \ln \left[\prod_{i=1}^n (1/p_i)^{p_i} \right] \quad (3.3)$$

Taking exponentials of both sides of Eq.(3.3) gives

$$\exp(S) = \prod_{i=1}^n (1/p_i)^{p_i} \quad (3.4)$$

The function $\exp(S)$ replaces S to give a modified maximum entropy problem (EM) in which the constraint (2.4) has taken a slightly different form.

(EM)

$$\begin{aligned} \text{Max} \quad G(p) &= \exp(S) = \prod_{i=1}^n (1/p_i)^{p_i} \\ \text{s.t.} \quad \sum_{i=1}^n p_i &= 1 \\ \sum_{i=1}^n p_i (g_j(x_i) - E[g_j]) &= 0 \end{aligned}$$

It can be noted here that taking the exponential of the entropy function S in no way affects its concavity and both problems (MEC) and (EM) have the same solution probabilities.

By comparing problem (EM) with the primal-dual pair of GP problems given in Appendix A, it can be immediately recognized that problem (EM) corresponds very closely to problem (DG), the GP dual problem.

Consequently, it is now possible to use the GP duality relationships as given in the Appendix to construct a new problem (DEM) which bears the same relationship to problem (EM) as (G) does to (DG).

(DEM)

$$\text{Min } F(y) = \sum_{i=1}^n \prod_{j=1}^m (y_j)^{g_j(x_i) - E[g_j]} \quad (3.5)$$

Thus we have established two pairs of primal-dual problems:

$$\text{(MEC) - (DE) and (EM) - (DEM)} \quad (3.6)$$

Since problem (EM) is obtained from (MEC) simply by taking the exponential of the entropy function S , it is expected that the two dual problems (DE) and (DEM) are closely related. In fact problem (DEM) can be obtained from problem (DE) by setting

$$y_j = \exp(\mu_j) \quad j=1, \dots, m \quad (3.7)$$

$$F(y) = \exp[D(\mu)] \quad (3.8)$$

It is therefore possible to calculate the least biased probabilities p_j by solving either of the dual problems, (DE) or (DEM).

3.3 INTERPRETATION OF THE ENTROPY DUALS

In interpreting the duality relationships obtained in the last section, it should be noted that the primal and dual problems are defined in different spatial domains. The primal problems (MEC) and (EM) are the well-known entropy maximization problems in the n -dimensional space of the discrete probabilities p_i whereas the dual problems (DE) and (DEM) are function minimization problems in the m -dimensional space of dual variables μ_j or y_j . This distinction is important in that the only point at which primal and dual spatial domains meet is at the unique optimizing point where entropy reaches its constrained maximum value while the dual function reaches its minimum value, and they are equal. There is, therefore, no transformation in a horizontal sense between a general, non-optimal point in primal space and another point in dual space. At a non-optimal point, primal and dual objective functions have different values, and the former is always less than the latter; that is,

$$S(p) = - \sum_{i=1}^n p_i \ln p_i \leq \ln Q - \sum_{j=1}^m \mu_j E[g_j] = D(\mu) \quad (3.9)$$

where the equality holds only at the respective optimum points, i.e., $S(p^*) = D(\mu^*)$.

There are many researchers who have missed this distinction between $S(p)$ and $D(\mu)$. For example, Tribus[97,98] has related the function

$D(\mu)$ to various thermodynamic functions without any reference to such a distinction. Jaynes was the first who indicated the importance of recognizing the difference between $S(p)$ and $D(\mu)$ in applying the maximum entropy criterion to irreversible processes. However, he did not appreciate the nature of the convex duality of $S(p)$ and $D(\mu)$, which the above work has demonstrated. He derived (3.9) by the ad hoc use of a mathematical inequality, and referred to the left-hand function and the right-hand function as informational entropy and experimental thermodynamic entropy, respectively. He presented this result in his Brandeis Lectures (See Ref.[63]). It is apparent that the relationship between these two entropies is shown more clearly through the present duality study than in Jaynes' work.

Geometric programming duality relationships have been derived by means of the arithmetic-geometric mean inequality as presented by Duffin et al[26]. The fact that problems (EM) and (DEM) form a primal-dual pair of a GP problem suggests that the dual formulation of an unconstrained GP problem represent an entropic process. This yields new insight into the dual formulation of GP problems. This duality relationship between GP problems and entropy maximization problems has been noted by several authors[23,104].

3.4 COMPUTATIONAL APPLICATION OF THE ENTROPY DUALS

A computationally effective algorithm for calculating least biased discrete probabilities p_i is very important in applying Jaynes' maximum entropy criterion. There have been several ways of calculating the desired probabilities. Griffeth[53] solved problem (MEC) directly using constrained nonlinear programming (a feasible direction method). An alternative method is to solve the m expectation equations for the m unknown Lagrange multipliers μ_j and substitute values of these multipliers into the n equations (2.9) to give the desired probabilities. The difficulty in this method lies in solving the m equations. One way of doing this is to use least squares methods by writing the equations in residual form. This is the basis of the methods used by Siddall and Diab[87], and Basu and Templeman[5] for continuous r.v.'s although they are equally applicable to discrete r.v.'s.

The above methods are cumbersome and require considerable care to be taken to avoid computational inaccuracies. A different approach, in the spirit of Agmon et al[1], is that of minimizing the dual function $D(\mu)$ directly and using Eq.(2.9) to calculate the desired probabilities. This approach is discussed further below.

Both dual problems (DE) and (DEM) are unconstrained nonlinear minimization problems. The number of dual variables is equal to the number of the expectation constraints m which is generally very small compared with the number of the primal variables n . Problem (DE) is here

chosen for study as the gradient of $D(\mu)$ has explicit expression obtained by substituting p_i from (2.9) into the expectation constraints (2.4). This means that efficient unconstrained minimization algorithms which employ gradients can be used. Another important feature is that it is always possible to specify a good starting point for the numerical search process. This consists of initially setting all the dual variables μ_j to zero. It is easily seen from Eq.(2.9) that this starting point is equivalent to making an initial guess that the probabilities p_i are uniformly distributed. If the numerical optimization process subsequently alters the values of the variables μ_j so as to minimize the dual function $D(\mu)$, this reflects the fact that, and the extent to which, the extra information provided by the statistical data causes the uniform probabilities to change to a different distribution. The strictly convex nature of problems (DE) and (DEM) ensures that the globally optimum point will be reached whatever the starting point may be.

A short program has been made to solve the dual problem (DE). A NAG library subroutine E04DFF, a modified Newton algorithm devised by Gill and Murray[45], is used for the unconstrained minimization of the function $D(\mu)$. A numerical example is calculated using this program and is presented next.

Example 3.1 The example is based upon an imaginary die-rolling experiment using an unfair die. The test No.1 postulates that after many rolls the mean of the face-up value is $E[x]=4.5$ (instead of 3.5 for a fair die). The corresponding problem (MEC) for this case consists of

maximizing an entropy function $S(p)$ over the variables $p_i (i=1, \dots, 6)$ subject to the normality condition (2.3) and one expectation constraint in the form

$$\sum_{i=1}^6 p_i \cdot i = 4.5$$

Table 3.1 Moment data used in test examples

Test No.	$E[x]$	$E[x^2]$	$E[x^3]$	$E[x^4]$
1	4.5	-	-	-
2	4.5	22.75	-	-
3	4.5	22.75	123.0	-
4	4.5	22.75	123.0	690.0

The dual function $D(\mu)$ involves only a single variable associated with the above constraint and is in the simplest form

$$D(\mu) = \ln \left[\sum_{i=1}^6 \exp(\mu i) \right] - 4.5\mu$$

which can be solved by any one-dimensional search and has the optimal solution $\mu^* = 0.371043$. Substituting this value into Eq.(2.9) gives the desired probabilities for this case, shown in Table 3.2.

In test No.2 it is assumed that both the mean and variance of the face-up values are known, $E[x]=4.5$ and $E[x^2]=22.75$ (instead of 15.666

for a fair die). In tests No.3 and No.4 hypothetical values for $E[x^3]$ and $E[x^4]$ are given as shown in Table 3.1. For each test the underlying discrete probabilities which maximize entropy subject to the corresponding expectation constraints are found using the short program. Results are given in Table 3.2.

Table 3.2 Resulting discrete probabilities

Test No.				
Probs.	1	2	3	4
p_1	.054354	.067414	.015168	.000030
p_2	.078772	.081046	.135295	.176936
p_3	.114161	.105414	.182140	.166814
p_4	.165447	.148336	.115131	.062205
p_5	.239774	.225829	.106299	.166814
p_6	.347491	.371961	.445967	.427054
S_{\max}	1.613581	1.609575	1.491612	1.440478

Computational experience shows that the dual algorithm needs little programming effort and converges very quickly. Numerical accuracy is very high.

3.5 DUAL FORMS FOR THE CONTINUOUS CASE

Dual forms similar in nature to (DE) and (DEM) can be derived for the continuous case in which the r.v. x is continuous-valued rather than discrete-valued. If $f(x)$ represents the probability density function of the continuous r.v. x , the maximum entropy problem (MECC) paralleling problem (MEC) is defined by (2.11)-(2.13).

Using the same approach as in Section 3.2 it can be shown that the dual forms are similar to problems (DE) and (DEM). One of them is derived using Lagrangean duality, and has the form:

(DEC)

$$\text{Min } D_c(\mu) = \ln \left[\int \exp \left(\sum_{j=1}^m \mu_j g_j(x) \right) dx \right] - \sum_{j=1}^m \mu_j E[g_j] \quad (3.10)$$

Another dual form parallels problem (DEM) and is in the form:

(DEMC)

$$\text{Min } F_c(y) = \int \prod_{j=1}^m (y_j)^{(g_j(x) - E[g_j])} dx$$

From a computational point of view, problems (DEC) and (DEMC) have all the attributes of their discrete counterparts. They are slightly more complicated in that they each involve a single integration to evaluate the dual function $D_c(\mu)$ or $F_c(y)$ at each trial point in the minimization process. However, the alternative solution methods for

continuous entropy maximization described in Refs.[5] and [87] require numerical integrations at each trial point. Consequently, obtaining maximum entropy p.d.'s via the dual forms (DEC) or (DEMC) should be considerably simpler and more rapid than the alternative methods.

3.6 CONCLUSIONS

As shown in the previous sections, the dual forms of the maximum entropy problem have merit in respect of both theoretical and computational aspects. They cast new insights into the nature of entropic processes and clarify some ambiguities in the literature. The distinction between informational entropy and experimental thermodynamic entropy has been made clear through the dual formulations and is very important in non-equilibrium thermodynamics. It has been demonstrated that the dual formulation for an unconstrained GP problem is an entropic process. This shows that the arithmetic-geometric mean inequality also involves entropic processes. This recognition make it possible to formulate the dual forms of GP problems through entropic processes.

The solution of the maximum entropy problem using the dual forms has shown considerable advantages over the primal approaches. This may promote more applications of Jaynes' maximum entropy criterion.

CHAPTER 4 ENTROPY IN OPTIMIZATION: DUAL METHODS

SUMMARY

This chapter studies the use of entropy in the solution of constrained nonlinear programming problems and is mainly concerned with developing dual solution methods. Through the surrogate Lagrangean, a constrained optimization problem is divided into a number of subproblems which are interpreted as the micro-states of a statistical thermodynamic system. Surrogate multipliers thus represent probabilities of the system being in each micro-state. An optimizing process is then compared with the transition of the system to an equilibrium state at which entropy of the system should attain a maximum value. Jaynes' maximum entropy criterion (MEC) is employed to formulate an entropy maximization model for calculating least biased probabilities, i.e., surrogate multipliers. A closed form of the solution is obtained which clearly demonstrates the interactivity among the multipliers. A primal-dual method is developed and a numerical example is presented.

The idea of entropy augmentation is then proposed in which a term related to the multiplier entropy is added to a surrogate dual. Using Lagrangean stationarity conditions for the augmented surrogate dual problem, we derive a dual iterative method which is proved to be equivalent to the entropy maximization method. However, if an explicit

surrogate dual can be obtained this approach requires iterations over dual space only rather than alternate iterations over primal and dual spaces as in the previous primal-dual method. This approach has certain similarities to the penalty function approach as well as the augmented Lagrangean approach.

4.1 INTRODUCTION

It has been pointed out in the end of Chapter 2 that the optimization problem to be studied in this thesis is the inequality constrained problem (P), as defined by (1.1) and (1.2). It is usually impossible for one to know in advance whether a particular constraint is active or not at the optimum, which has presented great difficulties in developing solution methods. Various active set strategies are usually employed to alleviate these difficulties. However, such a treatment destroys the relationships between the original constraints in the sense that even for a practical problem, it is possible for an active set to vary during the course of an optimization process. This creates discontinuities of some functions and their derivatives and causes some algorithms to suffer from instability. The present study aims at examining the "interactivity" among the constraints, hence among the Lagrange multipliers, and developing solution methods which are effective without relying upon any active set strategy.

The difficulty of making a prediction about the activity levels of the constraints arises from many possible combinations of active and inactive constraints. When there is a single constraint, there can exist only two states: either active or inactive. For a problem with m constraints, however, there are a total of 2^m different combinations. It is obvious that a solution method will not be efficient if it solves the original problem by solving 2^m problems each of which corresponds to a particular combination. This combinatorial complexity for a constrained optimization problem has some similarity to a statistical thermodynamic system. For example, we consider a one-dimensional lattice as shown below



Fig.4.1 One-Dimensional Lattice

At each point of the lattice, there is a particle whose spin direction is denoted by a small arrow, pointing up or down. If the lattice has m points, there are 2^m different possibilities of how to arrange the orientation of the arrows. Although it is difficult to predict the exact state of the system on the microscopic scale, the methods of statistical thermodynamics resolve this problem by estimating probabilities of the system being in each micro-state. Once these probabilities have been determined, some macroscopic thermodynamic quantities such as energy can be evaluated as the "mean" values of the corresponding thermodynamic quantities of individual micro-states. The combinatorial similarity

between a constrained optimization problem and a statistical thermodynamic system, such as the above one-dimensional lattice, suggests that the principles and methods of statistical thermodynamics could be used in the solution of constrained optimization problems.

4.2 STATISTICAL THERMODYNAMIC SIMULATION OF OPTIMIZATION

It seems to be natural that for an optimization problem with m inequality constraints, we divide it into m suboptimization problems (PJ), as defined by (2.26) and (2.27). Each of these subproblems is then interpreted as a micro-state of the simulated thermodynamic system which has a certain "energy" represented by the Lagrangean function for the corresponding subproblem (PJ):

$$L_j(x, \alpha_j) = f(x) + \alpha_j g_j(x) \quad (4.1)$$

where α_j is the Lagrange multiplier associated with a single constraint j . The system energy is then the mean value of energies L_j of all the individual micro-states in a probabilistic sense. Suppose that at a particular equilibrium state, probabilities p_j of each micro-state have been assigned, then the mean energy of the system is

$$L(x, p, \alpha) = \sum_{j=1}^m p_j L_j(x, \alpha_j) = \sum_{j=1}^m p_j [f(x) + \alpha_j g_j(x)] \quad (4.2)$$

or equivalently

$$L(x,p,\alpha) = f(x) + \sum_{j=1}^m p_j \alpha_j g_j(x) \quad (4.3)$$

because of the normalization of the probabilities p_j .

It should be noted that the mean energy function (4.3) has some disadvantages which prevent it from direct use in a thermodynamic system. Comparing Eq.(4.3) with the surrogate Lagrangean function (1.41) which is rewritten as follows

$$L(x,\lambda,\alpha) = f(x) + \alpha \sum_{j=1}^m \lambda_j g_j(x) \quad (1.41)$$

we find that the m individual Lagrange multipliers α_j in (4.3) are replaced by a single scalar multiplier α in (1.41) associated with the surrogate constraint. It is therefore convenient to use Eq.(1.41) rather than Eq.(4.3) as the mean energy of the system. Some relationships among the vectors p , λ and α , and the scalar α can be obtained by setting

$$\alpha \lambda_j = p_j \alpha_j \quad j=1, \dots, m \quad (4.4)$$

Since λ_j satisfy the normality condition (1.40), the summation of both sides of Eq.(4.4) over the index j from 1 to m yields

$$\alpha = \sum_{j=1}^m p_j \alpha_j \quad (4.5)$$

It turns out that the Lagrange multiplier α associated with the surrogate constraint is the mean value of the individual Lagrange multipliers α_j associated with each constraint. This has reflected the fact that the surrogate constraint represents a certain "average" of the original constraints. The relationships between the vectors λ and p can be derived by substituting α from (4.5) into (4.4) and are written as

$$\lambda_j = p_j \alpha_j / \sum_{j=1}^m p_j \alpha_j \quad j=1, \dots, m \quad (4.6)$$

which demonstrate a probabilistic interpretation for the surrogate multipliers λ_j .

Thus a slightly different statistical thermodynamic model is described as follows. For ease of explanation, the surrogate Lagrangean function (1.41) can be expressed in an equivalent form:

$$L(x, \lambda, \alpha) = \sum_{j=1}^m \lambda_j [f(x) + \alpha g_j(x)] \quad (4.7)$$

using the normalization of the surrogate multipliers λ_j . Then the simulated thermodynamic system can be thought as consisting of m micro-states, each of which has energy

$$L_j(x, \alpha) = f(x) + \alpha g_j(x) \quad (4.8)$$

The surrogate multipliers λ_j then represent probabilities of the system being in each micro-state.

Statistical thermodynamics tells us that a thermodynamic system spontaneously tends to its equilibrium state at which the probabilities within the system are given by the Boltzmann-Gibbs distribution. As demonstrated by Jaynes, the same distribution can be generated using the maximum entropy criterion. Thus an optimization problem is related to a thermodynamic equilibrium problem which reduces to an entropy maximization problem of calculating the least biased probabilities.

4.3 AN ENTROPY MAXIMIZATION APPROACH TO OPTIMIZATION

Although the surrogate constraint approach has the disadvantage of requiring a sequence of surrogate problems (S) to be solved, a surrogate constraint depicts the interactivity between original constraints better than in other approaches. In some cases, this approach has demonstrated its usefulness (Refs.[2], [24], [44], [48] and [49]). In Section 1.4 we have described a possible solution procedure and introduced the concept of surrogate duality. The major snag in a surrogate iterative scheme lies in updating the surrogate multipliers λ_j . It is clear from the surrogate duality theory that the updates

should maximize the surrogate dual function $s(\lambda)$ subject to the normality condition (1.40). However, there is generally no explicit dual function $s(\lambda)$. Surrogate algorithms have hitherto based the multiplier updates upon gradient information only which are similar to the Lagrange multiplier updates (1.34). It has been pointed out that these updates converge very slowly.

By analogy with thermodynamic equilibria, the updates of these multipliers are governed by entropic processes. Jaynes's maximum entropy criterion is then employed to construct the following entropy maximization problem:

(ME)

$$\text{Max} \quad H(\lambda) = - \sum_{j=1}^m \lambda_j \ln \lambda_j \quad (4.9)$$

$$\text{s.t.} \quad \sum_{j=1}^m \lambda_j = 1 \quad (1.40)$$

$$\sum_{j=1}^m \lambda_j g_j = e \quad (4.10)$$

where $H(\lambda)$ represents the entropy of the system, g_j represent values of the constraints $g_j(x)$ at the updating point and the scalar e corresponds to a change in the mean energy of the system when the system changes from one configuration to another as discussed in the next section.

Applying the Lagrange multiplier method to the above maximum entropy problem (ME) (as was done to problem (MEC) in Section 2.3), we have a closed form of the solution to problem (ME) as

$$\lambda_j = \exp(\beta g_j) / \sum_{j=1}^m \exp(\beta g_j) \quad j=1, \dots, m \quad (4.11)$$

where the scalar β represents the Lagrange multiplier associated with the constraint (4.10) and may be interpreted as a "temperature" parameter of the system, which is functionally related to the quantity e . Substituting (4.11) into (4.10) gives

$$\sum_{j=1}^m g_j \exp(\beta g_j) / \sum_{j=1}^m \exp(\beta g_j) = e \quad (4.12)$$

In the proposed algorithm, instead of specifying values for e and determining β through (4.12), we use the temperature parameter β as a control parameter whose value determines e .

Eq.(4.11) is the update formula for surrogate multipliers λ_j in a surrogate solution procedure. Taking a closer look at it, some features can be seen. Firstly, this formula has a single parameter β upon which all the surrogate multipliers λ_j depend. Thus the problem to update the m multipliers has become a much simpler single parameter problem. The nonnegativity and normality conditions required to be satisfied by the surrogate multipliers λ_j are automatically satisfied by Eq.(4.11) itself. Once β is determined, the updates for the surrogate multipliers

follow automatically from (4.11). Secondly, consider two particular surrogate multipliers λ_p and λ_q for instance. From (4.11), the ratio (λ_p/λ_q) is

$$\lambda_p/\lambda_q = \exp[\beta(g_p - g_q)] \quad (4.13)$$

which will be greater than unity if $g_p > g_q$ and $\beta > 0$. This accords with the intuition that the greater the activity level of a constraint, the greater its surrogate multiplier will be. This suggests that β should be positive. For the given values of the two constraints, the ratio λ_p/λ_q is determined by the magnitude of β , i.e., the greater β , the greater the updated surrogate multiplier associated with the constraint having the largest value will be. In other words, the constraint with the larger value will play a greater role in the next iteration for a surrogate constraint approach than it does for a smaller β . It is therefore sensible to specify an increasing positive sequence for the parameter β . Thus the effects of all of the constraints can be incorporated in the optimization process in the early stages and slack constraints can be discarded later as the solution of problem (S) approaches the solution of problem (P).

The solution procedure is then proposed as follows:

1. Set the iteration number $k=0$ and $\beta=0$, which amounts to setting all the surrogate multipliers to $\lambda_j=1/m$.
2. Solve problem (S) for given λ to yield the solution x^k .

3. Check some convergence criteria, for example the feasibility of the original constraints and the algorithm stops if they are satisfied.
4. Increase β by a positive value and update λ_j using (4.11).
5. Set $k=k+1$ and go to Step 2.

A small example is chosen to illustrate the above algorithm.

Example 4.1: The problem to be solved is chosen from Ref.[6] and takes the form

$$\begin{aligned}
 \text{Min} \quad & f(x) = (x_1 x_2 x_3)^{-1} \\
 \text{s.t.} \quad & g_1(x) = 2x_1 + x_2 + 3x_3 - 1 \leq 0 \\
 & g_2(x) = x_1 + x_2 + x_3 - 1 \leq 0 \\
 & g_3(x) = x_1 + 3x_2 + 2x_3 - 1 \leq 0
 \end{aligned}$$

The corresponding surrogate problem is:

$$\begin{aligned}
 \text{Min} \quad & f(x) = (x_1 x_2 x_3)^{-1} \\
 \text{s.t.} \quad & g_s(x) = \lambda_1 g_1(x) + \lambda_2 g_2(x) + \lambda_3 g_3(x) \leq 0
 \end{aligned}$$

The surrogate constraint can be rearranged on collecting the terms as

$$g_s(x) = (2\lambda_1 + \lambda_2 + \lambda_3)x_1 + (\lambda_1 + \lambda_2 + 3\lambda_3)x_2 + (3\lambda_1 + \lambda_2 + 2\lambda_3)x_3 - 1 \leq 0$$

This surrogate problem is a zero degree of difficulty GP problem and can be analytically solved for a given set of λ_j by solving a system

of linear equations (Ref.[26]). The solution is expressed in terms of the given multipliers λ_j ($j=1,2,3$) as

$$\begin{aligned}
 x_1 &= 1/[3(2\lambda_1+\lambda_2+\lambda_3)] \\
 x_2 &= 1/[3(\lambda_1+\lambda_2+3\lambda_3)] \\
 x_3 &= 1/[3(3\lambda_1+\lambda_2+2\lambda_3)]
 \end{aligned}$$

The parameter β is specified as $\exp(k)$ in this example where k represents the iteration number. The computational results are presented in Table 4.1. After 4 iterations, the optimal solution of the original problem is reached which is as exact as the three decimal place accuracy permits. Convergence is very fast.

Table 4.1 Results for Example 4.1

k	β	λ_1	λ_2	λ_3	x_1	x_2	x_3	g_1	g_2	g_3	f
0	-	1/3	1/3	1/3	.250	.200	.167	.200	-.383	.183	120.00
1	2.718	.463	.095	.442	.228	.177	.141	.056	-.454	.041	175.74
2	7.389	.521	.012	.467	.219	.172	.133	.009	-.476	.001	199.61
3	20.086	.540	.000	.460	.216	.174	.131	-.001	-.479	.000	203.11

4.4 AN ENTROPY AUGMENTATION APPROACH TO OPTIMIZATION

The entropy maximization approach presented in the last section is a primal-dual method which requires alternate iterations over both primal

and dual spaces. The iterations over the primal space involve solving a sequence of surrogate problems (S) while the iterations over the dual space involve the updates of dual variables λ which are given by the solution of problem (ME), i.e., by Eq.(4.11).

It is possible to replace the alternate iterations by iterations over a single space, either primal or dual space. This section examines a dual iterative method. The primal method is left to the next chapter.

A surrogate dual problem (SD) is defined by (1.44)-(1.46). In general, $s(\lambda)$ is an implicit function of the given surrogate multipliers λ except for some separable problems. This makes it difficult to solve the primal problem by means of solving the surrogate dual problem. Furthermore, even if an explicit dual function $s(\lambda)$ is obtainable the solution of problem (SD) is not as easy as it looks. Instead of solving problem (SD), we solve a modified problem (SDE) in which a term related to the multiplier entropy $H(\lambda)$, as defined by (4.9), is added to the surrogate dual $s(\lambda)$. The entropy augmented surrogate dual problem (SDE) has the form:

(SDE)

$$\text{Max} \quad s_p(\lambda) = s(\lambda) + H(\lambda)/p \quad (4.14)$$

$$\text{s.t.} \quad \sum_{j=1}^m \lambda_j - 1 = 0 \quad (4.15)$$

$$\lambda_j \geq 0 \quad j=1, \dots, m \quad (4.16)$$

where

$$H(\lambda) = - \sum_{j=1}^m \lambda_j \ln \lambda_j \quad (4.9)$$

and p is a parameter which is specified as an increasing positive sequence. The problem (SDE) with the parameter p is later referred to as the p th entropy augmented surrogate dual problem. It is intuitively obvious that as the parameter p becomes sufficiently large the term $s(\lambda)$ will dominate the solution to problem (SDE). Thus it seems reasonable to suppose that $s(\lambda)$ approaches the maximum of problem (SD) with p approaching infinity.

At first, let us examine some properties of problem (SDE). It is apparent that the solution of problem (SDE) is a function of the parameter p which is denoted by λ^P . Then the entropy augmented objective function is denoted by

$$s_p(\lambda^P) = s(\lambda^P) + H(\lambda^P)/p \quad (4.17)$$

The following theorem describes some of important properties of problem (SDE).

Theorem 4.1: Given $0 < p \leq q$, if λ^P and λ^Q are the solutions of the p th and q th problems (SDE), respectively, then

$$s(\lambda^P) \leq s(\lambda^Q) \quad (4.18)$$

$$H(\lambda^P) \geq H(\lambda^Q) \quad (4.19)$$

$$s_p(\lambda^P) \geq s_q(\lambda^Q) \quad (4.20)$$

Proof: since λ^p and λ^q are the solutions of problems (SDE) for given p and q , respectively, the following inequalities hold:

$$\begin{aligned} s_p(\lambda^q) &\leq s_p(\lambda^p) \\ s_q(\lambda^p) &\leq s_q(\lambda^q) \end{aligned}$$

that is,

$$s(\lambda^q) + H(\lambda^q)/p \leq s(\lambda^p) + H(\lambda^p)/p \quad (4.21)$$

$$s(\lambda^p) + H(\lambda^p)/q \leq s(\lambda^q) + H(\lambda^q)/q \quad (4.22)$$

Eliminating $H(\lambda^p)$ and $H(\lambda^q)$ from (4.21) and (4.22) gives

$$s(\lambda^p)(q - p) \leq s(\lambda^q)(q - p) \quad (4.23)$$

from which (4.18) follows. Similarly eliminating $s(\lambda^p)$ and $s(\lambda^q)$ gives

$$H(\lambda^q)(q - p) \leq H(\lambda^p)(q - p) \quad (4.24)$$

from which (4.19) follows. As $q \geq p$ and $H(\lambda) \geq 0$,

$$s(\lambda^q) + H(\lambda^q)/p \geq s(\lambda^q) + H(\lambda^q)/q = s_q(\lambda^q) \quad (4.25)$$

and λ_p is the solution of problem (SDE) for given p such that

$$s_p(\lambda^p) = s(\lambda^p) + H(\lambda^p)/p \geq s(\lambda^q) + H(\lambda^q)/p \quad (4.26)$$

from (4.25) and (4.26), (4.20) follows. This completes the proof.

===== ®

The theorem implies that $s(\lambda^P)$ will increase whereas $H(\lambda^P)$ and $s^P(\lambda_p)$ will decrease monotonically with increase of the parameter p . It is also assumed that the surrogate dual problem (SD) is well defined. This assumption and Theorem 4.1 therefore mean that the limits of $s(\lambda^P)$ and $H(\lambda^P)$ exist as p tends to infinity, and are, respectively, the maximum of $s(\lambda)$ and the minimum of $H(\lambda)$. These properties ensure that the solution of the p th problem (SDE) converges to the solution of problem (SD) as the parameter p approaches infinity.

The solution of the p th problem (SDE) can be found using the Lagrangean stationarity conditions. As the nonnegativity condition (4.16) is automatically satisfied by the solution itself, it will not be introduced into the Lagrangean function of problem (SDE) which thus takes the form:

$$L_p(\lambda, \mu) = s(\lambda) + H(\lambda)/p + \mu \left(\sum_{j=1}^m \lambda_j - 1 \right) \quad (4.27)$$

The stationarity conditions are

$$\partial L_p / \partial \lambda_j = \partial s(\lambda) / \partial \lambda_j + [\partial H(\lambda) / \partial \lambda_j] / p + \mu = 0 \quad j=1, \dots, m \quad (4.28)$$

where

$$\partial H(\lambda)/\partial \lambda_j = -(1 + \ln \lambda_j) \quad (4.29)$$

Substituting (4.29) into (4.28) and then rearranging (4.28) give

$$\lambda_j = \exp[p(\partial s(\lambda)/\partial \lambda_j + \mu) - 1] \quad (4.30)$$

The multiplier μ can be eliminated by considering the normality condition (4.15) and λ_j eventually takes the form

$$\lambda_j = \exp[p\partial s(\lambda)/\partial \lambda_j] / \sum_{j=1}^m \exp[p\partial s(\lambda)/\partial \lambda_j] \quad j=1, \dots, m \quad (4.31)$$

which is used as a recurrence formula for the surrogate multipliers λ_j .

For the case of an explicit $s(\lambda)$, its derivatives, $\partial s(\lambda)/\partial \lambda_j$, may also have explicit forms so that λ^P can be iteratively found using (4.31) and approaches λ^* as the parameter p tends to infinity.

A dual iterative solution procedure is as follows:

1. Set $p=0$, which amounts to setting all $\lambda_j=1/m$, and set an increment of p , δp .
2. Let $p=p+\delta p$ and iterate λ_j using (4.31) until the process converges to the solution λ^P . If the process converges very slowly or does not converge at all, set $\delta p=c\delta p$ ($c<1$) and repeat this step from the beginning.

3. Check global convergence criteria, and if satisfied the algorithm stops, otherwise return to step 2.

From Theorem 4.1, a global convergence criterion may check if $s(\lambda^p)$ remains increasing. Once $s(\lambda^p)$ starts dropping, the algorithm has converged. The local convergence criteria for Step 2 may check if specified errors for $s(\lambda)$ and $\|\lambda\|$ have been satisfied.

The proposed dual iterative method has remarkable advantages over other methods. First of all, the normality and nonnegativity conditions (4.15) and (4.16) are automatically satisfied by Eq.(4.31). Secondly, this method is extremely simple because of the fact that only algebraic operations are required to reach the optimum solution. Moreover, convergence will be assured if an appropriate positive sequence of p is specified.

It is shown in Chapter 6 that for a class of structural optimization problems, an explicit surrogate dual can be derived so that the present method can be effectively used for solving these optimum design problems.

4.5 EQUIVALENCE RELATIONSHIP BETWEEN EQS.(4.11) AND (4.31)

For the general case where an explicit surrogate dual $s(\lambda)$ cannot be found, Eq.(4.31) is still useful. The solution of problem (S) for a given vector λ should satisfy the following conditions:

$$\frac{\partial L_s}{\partial x_i} = \frac{\partial f(x)}{\partial x_i} + \alpha \sum_{j=1}^m \lambda_j \frac{\partial g_j(x)}{\partial x_i} = 0 \quad i=1, \dots, n \quad (4.32)$$

$$\frac{\partial L_s}{\partial \alpha} = \sum_{j=1}^m \lambda_j g_j(x) = 0 \quad (4.33)$$

The solution of problem (S) can then be obtained by solving Eqs.(4.32) and (4.33) which are expressed as $x(\lambda)$ and $\alpha(\lambda)$. Thus the surrogate dual function $s(\lambda)$ is an implicit function of λ obtained by substituting $x(\lambda)$ and $\alpha(\lambda)$ into the surrogate Lagrangean (1.41). Using the implicit function theorem and noting Eqs.(4.32) and (4.33), the derivatives of $s(\lambda)$ w.r.t. λ_j are calculated as:

$$\frac{\partial s(\lambda)}{\partial \lambda_j} = \alpha g_j(x) \quad j=1, \dots, m \quad (4.34)$$

Substituting (4.34) into (4.31) yields

$$\lambda_j = \frac{\exp[\rho \alpha g_j(x)]}{\sum_{j=1}^m \exp[\rho \alpha g_j(x)]} \quad j=1, \dots, m \quad (4.35)$$

If the product $p\alpha$ is replaced by a single parameter β , then Eq.(4.35) leads to the same updating formula as that given by Eq.(4.11), i.e., Eqs.(4.11) and (4.31) are equivalent. Consequently, we have proved that the entropy maximization approach is equivalent to the entropy augmentation approach. This justifies the previous entropy maximization approach as well as the statistical thermodynamic simulation.

4.6 CONCLUSIONS

Two entropy-based approaches to constrained optimization problems have been presented in the previous sections. By analogy with thermodynamic equilibria, Jaynes' maximum entropy criterion was employed to formulate an entropy maximization problem (ME) which yielded an update formula for surrogate multipliers. This formula is very simple and reflects the interactivity between the surrogate multipliers. The problem of estimating many multipliers has been reduced to a single parameter problem, which greatly simplifies the updating problem and strengthens the surrogate constraint approach. A numerical example was given which showed that the updates using Eq.(4.11) lead to fast convergence.

The idea of augmenting a surrogate dual $s(\lambda)$ by the multiplier entropy was presented, which has resulted in the derivation of a dual iterative method. This method requires iterations over only dual space if an explicit surrogate dual can be obtained. For the more general

case, the equivalence between the entropy maximization approach and the entropy augmentation approach was proved. This has provided additional insights into the entropy maximization approach as well as the statistical thermodynamic model.

In this chapter, the entropy augmentation approach was discussed in terms of the dual space to derive the dual iterative procedure. This approach has certain similarities to the penalty function approach. It can be also viewed as the augmented Lagrangean approach. This idea will be used in the next chapter to derive primal methods for constrained optimization.

CHAPTER 5 ENTROPY IN OPTIMIZATION: PRIMAL METHODS

SUMMARY

This chapter develops further the idea of entropy augmentation, as presented in the last chapter, to derive a penalty-like function $\Gamma_p(x, \alpha)$ for general constrained optimization which is referred to as the primal function of constrained optimization. The solution methods developed on the basis of this function are then referred to as the primal methods of constrained optimization.

Some important properties of the primal function are further examined which either supply theoretical interpretations for the primal function or prove to be useful in developing numerical algorithms. It is essentially proved that the function $\Gamma_p(x, \alpha)$ is a natural logarithm of the pth norm of a vector which consists of components $L_j(x, \alpha)$, the energy functions of the simulated thermodynamic system introduced in the last chapter. The present approach in fact converts a constrained optimization problem into a minimax optimization problem. The exponential-logarithmic transformations contained in the course of deriving the primal function make it easy to extend the present approach to a wide range of minimax optimization problems such as multicriteria optimization, curve fitting, etc.

There are many possibilities to use this function in the solution of optimization problems. Nice differentiability and monotonically decreasing properties ensure that many standard unconstrained minimization algorithms can be effectively applied to it and will converge to the solution of the original problem. An algorithm of the exact penalty function type is proposed and an example problem is presented.

5.1 INTRODUCTION

The entropy augmentation approach proposed in Chapter 4 is set in dual space and has demonstrated its usefulness both in deriving the dual iterative method and in justifying the entropy maximization approach as well as the statistical thermodynamic simulation of optimization. As noted there, however, the dual iterative method can be used only in the case where an explicit dual $s(\lambda)$ can be obtained, and the primal-dual method based on the entropy maximization approach is essentially oriented towards providing efficient updates of surrogate multipliers and relies upon efficient algorithms to solve a sequence of problems (S).

This chapter aims at further exploring implications of the entropy augmentation approach by developing solution methods for general constrained optimization problems. Here the framework of the surrogate

constraint approach remains to be used. To facilitate later presentations, we rewrite the surrogate Lagrangean as follows:

$$L_s(x, \lambda, \alpha) = f(x) + \alpha \sum_{j=1}^m \lambda_j g_j(x) \quad (1.41)$$

It has been noted in Chapter 4 that the surrogate dual function $s(\lambda)$ is obtained by first solving a surrogate problem (S) or, equivalently, by first minimizing the surrogate Lagrangean (1.41) with given λ over x -space for a solution $x(\lambda)$, and then substituting the minimizer $x(\lambda)$ into the surrogate Lagrangean. The solution of problem (P) is then obtained by maximizing the derived surrogate dual $s(\lambda)$ over λ -space which may be explicit or implicit. In this chapter we invert this process by first performing a maximization of the surrogate Lagrangean (1.41) over λ -space and then substituting the maximizer $\lambda(x, \alpha)$ into the surrogate Lagrangean to obtain an explicit function of variables x and α . The resulting function is then minimized over x - α -space to obtain the solution of problem (P).

It is unfortunate that the desired explicit function cannot be obtained by directly maximizing the surrogate Lagrangean (1.41) as it is impossible to find an explicit maximizer $\lambda(x, \alpha)$. However, an explicit maximizer $\lambda(x, \alpha)$ can be obtained by maximizing a modified surrogate Lagrangean, i.e., the entropy augmented surrogate Lagrangean. The procedure follows that for deriving Eq.(4.31). Since an explicit surrogate dual $s(\lambda)$ cannot be obtained, we replace $s(\lambda)$ in the entropy augmented dual Lagrangean (4.27) by the surrogate Lagrangean such that

the entropy augmented Lagrangean (4.27) is expressed in terms of both primal and dual spaces and has an alternative form

$$L_p(x, \lambda, \alpha, \mu) = f(x) + \alpha \sum_{j=1}^m \lambda_j g_j(x) - (1/p) \sum_{j=1}^m \lambda_j \ln \lambda_j + \mu (\sum_{j=1}^m \lambda_j - 1) \quad (5.1)$$

The next section is devoted to deriving the desired function by maximizing (5.1) over λ -space and to presenting some interpretations for this function.

5.2 PRIMAL FUNCTION OF CONSTRAINED OPTIMIZATION

It can be readily verified by examining the second derivatives of (5.1) w.r.t λ_j that the entropy augmented surrogate Lagrangean is a concave function of the variables λ_j . Its maximization over λ -space can therefore be carried out using the Lagrangean stationarity conditions w.r.t. λ . The procedure is exactly the same as that for deriving Eq.(4.31). As each stationarity equation involves a single surrogate multiplier, the explicit solution $\lambda(x, \alpha)$ can be easily found as

$$\lambda_j(x, \alpha) = \exp[\alpha g_j(x)] / \sum_{j=1}^m \exp[\alpha g_j(x)] \quad j=1, \dots, m \quad (5.2)$$

where p is regarded as a control parameter. It can be recognized that the maximizing variables λ_j have the same analytical expressions as those given by Eq.(4.35) as a consequence of having used the implicit function theorem in deriving (4.35). Substituting the explicit solution $\lambda(x,\alpha)$ into (5.1) yields

$$\Gamma_p(x,\alpha) = f(x) + (1/p)\ln \sum_{j=1}^m \exp[p\alpha g_j(x)] \quad (5.3)$$

which is the desired explicit function and is, throughout this thesis, referred to as the primal function of the constrained optimization problem (P).

An alternative form of (5.3) can be obtained by replacing the product $p\alpha$ with a parameter r and has the form:

$$\Gamma_r(x,\alpha) = f(x) + \alpha G_r(x) \quad (5.4)$$

where

$$G_r(x) = (1/r)\ln \sum_{j=1}^m \exp[rg_j(x)] \quad (5.5)$$

The primal function Γ_p or G_r may also be called the potential function of constrained optimization. The first term, the objective function $f(x)$, may be thought of representing the internal potential "energy" of the system while the second term represents the external potential

or constraint potential "energy" of the system which incorporates the effects of all the constraints on the optimization.

It should be emphasized that the key to deriving the primal function Γ_p or Γ_r explicitly expressed in terms x and α consists in the ability to find an explicit maximizer $\lambda(x, \alpha)$ of the entropy augmented Lagrangean (5.1) over λ -space. This approach becomes possible if and only if the surrogate Lagrangean (1.41) has been augmented by the multiplier entropy term $H(\lambda)/p$. At this point, the distinction between the entropy augmented Lagrangean (5.1) and the surrogate Lagrangean (1.41) is made clear.

Some implications of the function $G_r(x)$ are examined next by comparing the surrogate Lagrangean (1.41) with the entropy augmented surrogate Lagrangean (5.1). Collecting the second term and the third term of (5.1) together yields

$$\alpha \sum_{j=1}^m \lambda_j [g_j(x) - (1/p\alpha) \ln \lambda_j] \quad (5.6)$$

from which it is easy to recognize that each original constraint g_j in (1.41) has been replaced in (5.1) by

$$g_j' = g_j(x) - (1/p\alpha) \ln \lambda_j \quad j=1, \dots, m \quad (5.7)$$

in which the second term is an additional term and is always nonnegative since $p > 0$ and $\lambda_j \leq 1$. Therefore, if the original constraint inequality

$g_j \leq 0$ is replaced by $g_j' \leq 0$, each constraint becomes tighter than its unpenalized counterpart $g_j \leq 0$. This effect represents the penalty upon each constraint, i.e., upon an optimization process. This type of penalty, referred later to as an entropy penalty, is completely different from conventional penalty approaches in which the penalties are imposed individually whereas the entropy penalty is imposed integratedly and consistently. In other words, the entropy penalty is imposed upon the whole constraint set rather than upon the individual constraints. This can be seen more clearly if the explicit solution $\lambda_j(x, \alpha)$ of (5.2) is substituted into (5.7), which leads to

$$g_j' = (1/p\alpha) \ln \sum_{j=1}^m \exp[p\alpha g_j(x)] \quad j=1, \dots, m \quad (5.8a)$$

or its equivalent form:

$$g_j' = (1/r) \ln \sum_{j=1}^m \exp[rg_j(x)] \quad j=1, \dots, m \quad (5.8b)$$

where the entropy penalized constraints g_j' have been expressed in terms of the primal variables. A remarkable thing happens in that the constraints g_j' which appear to be different from each other in terms of the variables λ , x and α , as shown in Eq.(5.7), become all identical and equal to the function $G_r(x)$. This implies that the individual constraints are aggregated into and replaced by a single constraint

$$G_r(x) \leq 0 \quad (5.9)$$

where r is a control parameter. This result can be obtained only in the case where the λ_j take the values (5.2) obtained by the entropy augmentation approach, i.e., the entropy penalty approach. It is then argued that the entropy penalty is the most consistent as it reflects the integrity and the interactivity of all the constraints and thus makes the obtained penalty function continuous and differentiable.

If $G_r(x)$ is regarded as a single parametric constraint, then the primal function $\Gamma_r(x, \alpha)$ of (5.4) can also be interpreted as the Lagrangean function of a constrained optimization problem

(PP)

$$\begin{array}{ll} \text{Min} & f(x) \\ \text{s.t.} & G_r(x) \leq 0 \end{array} \quad (5.9)$$

where $G_r(x)$ is given by (5.5). Instead of optimizing $\Gamma_p(x, \alpha)$ over both x and α , one may find it advantageous to solve problem (PP). It is interesting to make a comparison of the parametric constraint $G_r(x)$ with the original constraint set. The feasible regions of this parametric constraint for different values of the parameter r are depicted in Fig.5.1 and compared with the primal feasible region. It can be seen that the smaller the parameter r , the smaller the feasible region of the parametric constraint $G_r(x)$ will be. Theoretically, the primal constraint set corresponds to a parametric constraint in which the parameter p approaches infinity. However, the primal constraint set can be satisfactorily simulated by a moderate parameter r , as shown in Fig.5.1. It is then possible to use a moderately large parameter r to

solve a constrained optimization problem for high accuracy. This has been supported by the computation of Example 5.1 presented later. This property is similar to that of the multiplier methods introduced in Section 1.3.4.

5.3 SOME IMPORTANT PROPERTIES OF THE PRIMAL FUNCTION

Some important properties of the primal function $\Gamma_p(x, \alpha)$ are studied in this section. There are two main objectives in carrying out this study. The first is to further explore implications of the primal function Γ_p and the second is to discover some links between the present approach and other well-established mathematical results.

In Chapter 3, we have found that the dual formulation of an unconstrained GP problem is an entropic process. However the GP duality theory is based upon the arithmetic-geometric mean inequality (A.1) introduced in Appendix A. Moreover, the weights in the inequality play a role similar to that of the surrogate multipliers. It is consequently conjectured that an investigation into the relationship between the inequality and the entropy augmentation approach should supply additional insights into the results we have obtained. The main results from this study are presented in Theorem 5.1. The arguments of some functions are omitted in the following presentations for the sake of clarity.

Theorem 5.1: The primal function Γ_p is an upper bound on both the entropy augmented surrogate Lagrangean L_p of (5.1) and the surrogate Lagrangean L_s of (1.41), i.e., $\Gamma_p \geq L_p \geq L_s$ holds for all λ .

Proof: To prove the above relation, the arithmetic-geometric mean inequality (A.1) is employed and rewritten as follows

$$\sum_{j=1}^m u_j \geq \prod_{j=1}^m (u_j/\lambda_j)^{\lambda_j} \quad (\text{A.1})$$

where all u_j are nonnegative, and λ_j are positive weights satisfying the following normality condition

$$\sum_{j=1}^m \lambda_j = 1 \quad (\text{A.2})$$

The inequality (A.1) becomes an equality if and only if all u_j are equal.

In Section 4.2, we have divided a constrained optimization problem (P) into m subproblems each of which has a "energy" function

$$L_j(x, \alpha) = f(x) + \alpha g_j(x) \quad j=1, \dots, m \quad (4.8)$$

which may have positive or negative value. However, by setting

$$u_j = [\exp(L_j)]^P \quad j=1, \dots, m \quad (5.10)$$

where $p > 0$, we ensure strictly positive components u_j ($j=1, \dots, m$). Thus the inequality (A.1) can be used. Substituting u_j from (5.10) into (A.1) gives

$$\sum_{j=1}^m [\exp(L_j)]^p \geq \prod_{j=1}^m \{[\exp(L_j)]^p / \lambda_j\}^{\lambda_j} \quad (5.11)$$

Raising (5.11) to the power $(1/p)$ and then taking natural logarithms of both sides of the inequality, after some manipulation we have an inequality

$$(1/p) \ln \sum_{j=1}^m \exp(pL_j) \geq \sum_{j=1}^m \lambda_j L_j - (1/p) \sum_{j=1}^m \lambda_j \ln \lambda_j \quad (5.12)$$

It turns out that by substituting L_j from (4.8) into (5.12) the left-hand side becomes the primal function $\Gamma_p(x, \alpha)$ while the right-hand side is a variant of the entropy augmented surrogate Lagrangean (5.1). We then have

$$\Gamma_p(x, \alpha) \geq L_p'(x, \lambda, \alpha) \quad (5.13)$$

where $L_p'(x, \lambda, \alpha)$ is given by

$$L_p'(x, \lambda, \alpha) = f(x) + \alpha \sum_{j=1}^m \lambda_j g_j(x) - (1/p) \sum_{j=1}^m \lambda_j \ln \lambda_j \quad (5.14)$$

in which the last term of (5.1) related to the normality condition has been removed due to the normalization of the weights λ_j . Thus L_p of (5.1) is equivalent to L_p' of (5.14), and may replace L_p' in (5.13). As the inequality (5.13) holds regardless of values of the weights λ_j , $\Gamma_p \geq L_p' = L_p$ holds for all λ satisfying the normality condition. This completes the first part of the proof.

The proof of the second part is trivial. As the additional entropy term in L_p' is always positive, $L_p = L_p' \geq L_s$ simply follows. Together with (5.13), we have $\Gamma_p \geq L_p \geq L_s$. This completes the proof.

®

From the proof procedure of Theorem (5.1), there are several important points to be noted.

1. An "upper bound" means that for the same x and α , the primal function Γ_p is always greater than both L_p and L_s for all λ . This conclusion, drawn from the arithmetic-geometric mean inequality, confirms and verifies the fact that the function Γ_p may be obtained by the maximization of L_p over λ -space as was shown earlier. Thus the two different approaches have led to the same conclusion. This shows links between entropic processes and the arithmetic-geometric mean inequality as well as GP dualities. From a viewpoint of minimax optimization, the maximization phase has been realized by the maximization process over λ . The minimization phase requires the function Γ_p to be minimized over primal space.

2. It can be noted that the primal function Γ_p is the natural logarithm of the pth norm of a vector, i.e.,

$$\Gamma_p = \ln \Omega_p \quad (5.15)$$

where

$$\Omega_p = \left\{ \sum_{j=1}^m [\exp(L_j)]^p \right\}^{(1/p)} \quad (5.16)$$

The vector, denoted by Ψ , consists of components

$$\Psi_j = \exp(L_j) = \exp[f(x) + \alpha g_j(x)] \quad j=1, \dots, m \quad (5.17)$$

which are the exponentials of "energy" functions (4.8) of the simulated thermodynamic system introduced in the last chapter. Therefore, minimizing Γ_p amounts to minimizing the pth norm of this vector, and has the limit $\text{Min}[\max(L_j)]$ as $p \rightarrow \infty$, i.e., minimizing the maximum component of the energy functions $L_j(x, \alpha)$ of the thermodynamic system. Then the present approach is essentially a minimax optimization approach. It is similar to the so-called least pth optimization methods (Refs. [4], [19-21]). However, the exponential transformations contained in the present approach ensure strictly positive components and enable the maximum component to be approached quickly. This property may suggest that a small value of parameter p should be sufficient to obtain the solution to the original problem.

The following theorem shows a nice decreasing property of the primal function $\Gamma_p(x, \alpha)$ which is very useful in developing numerical algorithms. It is further proved that the primal function Γ_p approaches the maximum component of the energy functions (4.8) as p tends to infinity. Theorem 5.2 thus strengthens and refines the conclusions of Theorem 5.1.

Theorem 5.2: If $0 < q \leq p$, then $\Gamma_p \leq \Gamma_q$ holds for all x and α . Furthermore, Γ_p approaches $\max(L_j)$ as p tends to infinity.

Proof: According to Jensen's inequality[57], if $p \geq q$, then the p th norm of a vector is always less than or equal to its q th norm. From Eq.(5.16), $\Omega_p \leq \Omega_q$ simply follows if $p \geq q$. As a natural logarithmic function increases monotonically with increase in its argument, it readily follows that

$$\Gamma_p(x, \alpha) \leq \Gamma_q(x, \alpha) \quad \text{if } p \geq q. \quad (5.18)$$

which means that the primal function $\Gamma_p(x, \alpha)$ decreases monotonically as p increases. From the definition of the uniform (∞) norm, we have

$$\lim_{p \rightarrow \infty} \Omega_p = \max[\Psi_j] = \max[\exp(L_j)] \quad (5.19)$$

so that

$$\lim_{p \rightarrow \infty} \Gamma_p = \lim_{p \rightarrow \infty} \ln \Omega_p = \ln \lim_{p \rightarrow \infty} \Omega_p = \ln[\max(\exp(L_j))] = \max(L_j) \quad (5.20)$$

where we have used the continuity property of the natural logarithmic function which permits the two operations \ln and \lim to be exchanged. Eq.(5.20) means that Γ_p approaches the maximum component of the energy functions as p tends to infinity. This completes the proof.

®

From Theorem 5.1 and 5.2, some important properties of the function $\Gamma_p(x, \alpha)$ are summarized as follows:

1. $\Gamma_p \geq L_p \geq L_s$ for all $\lambda \geq 0$ where L_p and L_s are the entropy augmented and unaugmented surrogate Lagrangeans, respectively.
2. Γ_p may be regarded as a special p th norm, termed the p th logarithmic norm, as it can be obtained by taking the logarithm of the p th norm Ω_p of a vector Ψ .
3. Γ_p monotonically decreases as p increases, and it approaches $\max(L_j) = \max(f + \alpha g_j)$ as $p \rightarrow \infty$. This then implies

$$\lim_{p \rightarrow \infty} \text{Min}[\Gamma_p] = \text{Min}[\lim_{p \rightarrow \infty} \Gamma_p] = \text{Min}[\max(L_j)] = \text{Min}[\max(f + \alpha g_j)] \quad (5.21)$$

which means that solving problem (P) by means of solving a sequence of minimizations of Γ_p with increasing values of p is equivalent to transforming problem (P) into a minimax optimization problem.

4. Γ_p is continuous and differentiable as many times as are problem functions $f(x)$ and $g_j(x)$ ($j=1, \dots, m$).

When devising numerical algorithms using $\Gamma_p(x, \alpha)$, these properties ensure that efficient unconstrained minimization algorithms can be used and stable convergence can be expected.

5.4 DEVELOPMENT OF NUMERICAL ALGORITHMS

There are several ways of using the primal function $\Gamma_p(x, \alpha)$ or $\Gamma_r(x, \alpha)$ to develop numerical algorithms for solving the inequality constrained problem (P). They are briefly outlined below.

1. Primal approach: Primal methods such as the feasible direction methods and the gradient projection methods introduced in Section 1.3.1 can be applied to the single constrained problem (PP) defined in Section 5.2. The algorithms may start from a feasible point and approach the boundaries of feasible regions as the parameter r increases. They are therefore similar to the barrier function methods introduced in Section 1.3.2.
2. Sequential unconstrained minimization approach: A sequence of the p th problems

(Pth)

$$\text{Min } \Gamma_p(x, \alpha) = f(x) + (1/p) \ln \sum_{j=1}^m \exp[p\alpha g_j(x)] \quad (5.3)$$

is solved to approach the solution of problem (P) as $p \rightarrow \infty$. In each p th minimization, the solutions x^p and α^p are obtained. It is expected that this sequence converges to the solution of problem (P). The procedure is as follows. An initial parameter p is specified. Problem (Pth) is then solved by an unconstrained minimization algorithm to obtain the solution x^p and α^p . They are then used as the initial points of the next minimization. The parameter p is increased by an increment δp which does not have to be the same value in each minimization. Problem (Pth) with the parameter $p + \delta p$ is solved again. This process is repeated until convergence is reached.

3. Exact penalty function approach: The function Γ_p is regarded as an exact penalty function so that only one unconstrained minimization is required to solve problem (P). This approach can also be viewed as a simplified version of the sequential minimization approach; that is, for a particular parameter p , problem (Pth) does not have to be completely solved. It is recommended that only one minimization step be taken for a particular p . Although this approach needs the parameter p to be frequently increased, the total computational effort can be considerably reduced.

It should be noted that gradient-based algorithms can be easily adapted to the minimizations of the function $\Gamma_p(x, \alpha)$ because the first derivatives of $\Gamma_p(x, \alpha)$ w.r.t. x and α can be easily evaluated by

$$\frac{\partial \Gamma_p}{\partial x_i} = \frac{\partial f}{\partial x_i} + \alpha \sum_{j=1}^m \lambda_j (\frac{\partial g_j}{\partial x_i}) \quad i=1, \dots, n \quad (5.22)$$

and

$$\frac{\partial \Gamma_p}{\partial \alpha} = \sum_{j=1}^m \lambda_j g_j \quad (5.23)$$

where λ_j ($j=1, \dots, m$) are defined by Eq.(5.2). Another attractive feature of using the primal function $\Gamma_p(x, \alpha)$ to develop numerical algorithms is that as this function and its derivatives are expressed in simple and explicit forms, little programming effort is needed to incorporate the methods suggested above in a practical computer program. A basic numerical algorithm is developed to illustrate the use of the primal function $\Gamma_p(x, \alpha)$ in the solution of a constrained optimization problem, which is an exact penalty function algorithm. The algorithm is used to solve an example problem and is as follows:

1. Set the iteration number $k=0$. Specify p^0 and x^0 . α^0 is then estimated to make the residuals of the stationarity equations w.r.t. x small.
2. Set $k=k+1$ and use the negative gradient direction, i.e., $-\nabla_x \Gamma_p$ and $(\partial \Gamma_p / \partial \alpha)$, as the search direction.
3. Move along the above direction by a fixed small step-size to the point x^k and α^k .
4. Check convergence criteria, if satisfied, then stop.
5. Set $p=p+\delta p$, go to Step 2.

In the present algorithm, α and x are given an equal status. This makes the algorithm converge stably. An example is calculated by this algorithm and presented below.

Example 5.1: The problem to be solved is

$$\begin{aligned} \text{Min} \quad & f(x) = (x_1 - 2)^2 + (x_2 - 1)^2 \\ \text{s.t.} \quad & g_1(x) = x_1^2 - x_2 \leq 0 \\ & g_2(x) = x_1 + x_2 - 2 \leq 0 \end{aligned}$$

The point (2.0, 2.0) is chosen as a starting point x^0 . α^0 is then estimated as $\alpha^0 = 0.0$. The parameter p starts from 0.0 and increases by 0.1 at a time. The algorithm proceeds along the steepest descent directions by a fixed step-size of 0.1. In 20 steps, the solution point is near the optimum with the constraint g_1 being slightly violated. After 22 steps, the two constraints are all satisfied, but a specified accuracy 10^{-4} is not reached. The solution continues to be improved and converges after 60 steps. The results are listed in Table 5.1. It can be seen from Table 5.1 that the solution point approaches the optimum very fast within the first 20 steps, and improves marginally afterwards. There are two possible ways in which faster rates of convergence may be achieved. One of them is to use conjugate gradient or quasi-Newton directions as the search directions instead of the steepest descent directions. The second involves a linear search along each search direction.

Table 5.1 Results of Example 5.1

No	p	α	x_1	x_2	g_1	g_2	f	Γ_p
0	0.1	0.000	2.000	2.000	2.000	2.000	1.000	-
5	0.5	0.897	1.620	1.336	0.956	1.288	0.257	2.656
10	1.0	1.252	1.175	1.118	0.293	0.262	0.695	1.736
15	1.5	1.335	1.038	1.029	0.067	0.049	0.926	1.466
20	2.0	1.350	1.002	1.004	0.006	0.000	0.996	1.347
25	2.5	1.349	0.995	0.999	-0.007	-0.009	1.010	1.277
30	3.0	1.345	0.995	0.998	-0.007	-0.008	1.010	1.231
35	3.5	1.341	0.996	0.998	-0.006	-0.006	1.008	1.198
40	4.0	1.339	0.997	0.999	-0.004	-0.004	1.005	1.173
45	4.5	1.337	0.998	0.999	-0.003	-0.003	1.004	1.154
50	5.0	1.336	0.999	0.999	-0.002	-0.002	1.002	1.139
55	5.5	1.335	0.999	1.000	-0.001	-0.001	1.002	1.126
60	6.0	1.334	1.000	1.000	-0.001	-0.001	1.001	1.118

5.5 DISCUSSIONS

The function $G_r(x)$ has appeared in the literature (Refs.[56], [66], [89]) It was referred to in the paper by Kreisselmeier and Steinhauser[66]. They used the function as an approximation to the largest component of a vector, but gave no formal justification for it. Hajela[56] and Sobieski[89] used it without further examination in the solution of structural optimization problems to replace the original constraint set, and called it a cumulative constraint function.

In Section 5.2, we have formally derived this function $G_r(x)$ through the entropy augmentation approach. The derivation given there could

serve as a formal justification for its uses as a cumulative constraint and as an approximation to the largest component of a vector in a vector optimization context.

The present approach to solving problem (P) has some similarity to the least pth optimization methods proposed by Charalambous et al(Refs.[4], [19-21]) where problem (P) is transformed into a sequence of minimizations of the pth norm of a vector which consists of m+1 components

$$f_0 = f - f_{\text{opt}} \quad (5.24)$$

$$f_j = f - f_{\text{opt}} + \alpha_j g_j \quad j=1, \dots, m \quad (5.25)$$

where f_{opt} represents the optimum value of the objective function and α_j ($j=1, \dots, m$) are the multipliers associated with the constraints. They all need to be updated in the minimization of each pth problem.

In the present approach, the m multipliers α_j have been reduced to a single multiplier α so that the update of α becomes very simple and is made on an equivalent status level as the primal variables x_i ($i=1, \dots, n$), as done in the algorithm proposed in Section 5.4. Moreover, the exponential transformations contained in the present method have eliminated the requirement for the parameter f_{opt} to ensure positive values for each component. There are therefore considerable savings in computational and programming efforts.

The present approach can be easily adapted to solve general minimax optimization problems. If a minimax problem is posed as

$$\text{Min}[\max(f_j)] \quad (5.26)$$

then the minimax Lagrangean can be constructed as

$$L(x, \lambda) = \sum_{j=1}^m \lambda_j f_j \quad (5.27)$$

where λ_j are the Lagrange multipliers which satisfy the normality condition. By adding an entropy term $H(\lambda)/p$ to Eq.(5.27) and minimizing the entropy augmented minimax Lagrangean, we obtain an explicit maximizer

$$\lambda_j = \exp(pf_j) / \sum_{j=1}^m \exp(pf_j) \quad j=1, \dots, m \quad (5.28)$$

Substituting λ_j from (5.28) into (5.27), we have

$$F_p = (1/p) \ln \sum_{j=1}^m \exp(pf_j) \quad (5.29)$$

which may be referred to as the minimax primal function. The solution of the minimax optimization problem can therefore be obtained by solving a sequence of p th minimization problems in which the minimax primal

function F_p is minimized over primal variables for an increasing positive sequence of the parameter p .

5.6 CONCLUSIONS

In the previous sections, we have derived the primal function $\Gamma_p(x, \alpha)$ and examined some of its properties. It is essentially proved that this function is a natural logarithm of the p th norm of a vector, and thus the present approach transforms problem (P) into a minimax optimization problem. The vector consists of the exponentials of the energy functions $L_j = f + \alpha g_j$ of the simulated thermodynamic system. This justifies the division of problem (P) into m subproblems as well as the thermodynamic simulation of constrained optimization.

In the last section we have discussed how to use the present approach in solving general minimax optimization problems such as multicriteria optimization and curve fitting problems, etc. It has been shown that the approach presented herein handles this type of minimax problems quite easily.

Several ways of using this function in the solution of constrained optimization problems have been outlined. Its differentiability and monotonically decreasing properties make the function easy to use in numerical algorithms. An exact penalty function algorithm has been

developed to illustrate its use and an example problem has been presented. The present algorithm found a near-optimal solution quickly. However, as the solution point approached the optimum, more iterations were required to obtain an accurate solution. More computational experiences are needed to enable the primal function to be used extensively.

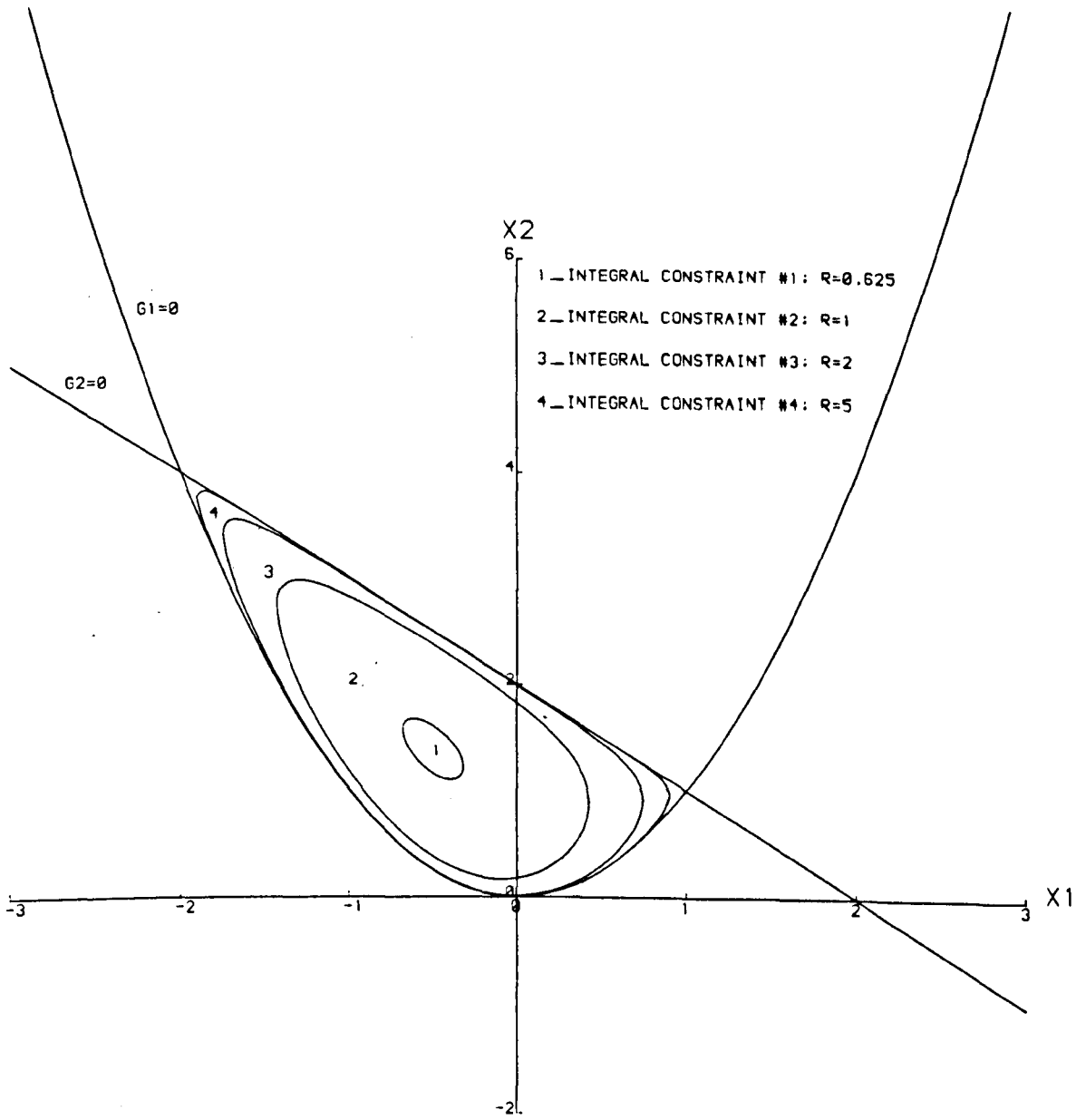


Fig.5.1 Comparisons between primal and integrated feasible regions

CHAPTER 6 ENTROPY METHODS FOR STRUCTURAL OPTIMIZATION

SUMMARY

This chapter is concerned with the solution of a class of structural optimization problems which is restricted to the weight minimization of a finite element model with fixed geometry and material properties. Under some assumptions and with some simplifications, these structural optimization problems can be formulated as a separable convex programming problem so that the corresponding surrogate problem (S) can be analytically solved to obtain the solution in terms of surrogate multipliers. Thus, the whole problem hinges upon an appropriate update scheme for these multipliers. As presented in Chapter 4, the update formula (4.11) developed by the entropy maximization approach therefore fills in this need. A primal-dual algorithm is then developed for solving this class of structural optimization problems in which the entropy-based updates are combined with the analytical solutions of surrogate problems.

The most important feature of this algorithm is that it does not require an active/passive set strategy, as most other algorithms do, so that all the constraints can be treated consistently throughout the optimization process. Moreover, the algorithm is so simple that only some algebraic operations are needed to produce an optimum design and,

therefore, both the number of design variables and the number of constraints do not affect the computational efforts as dramatically as is the case with other algorithms. The present algorithm is thus very suitable for solving the optimization problems of large-scale structures.

Several example problems are computed by the present algorithm and the results are compared with those previously reported. Computational experience shows that the performance of the entropy-based primal-dual algorithm is very satisfactory.

An explicit surrogate dual for the class of structural optimization problems is then derived using both Lagrangean stationarity conditions and the geometric programming duality theory. It turns out that the derived dual problem is exactly the same as that proposed by Templeman[91]. The derivations presented here have demonstrated that the dual formulation bears a surrogate duality relationship between the primal and the dual problems. Furthermore, as the dual function has an explicit form, the dual solution method developed by the idea of entropy augmentation can be used to effectively solve the surrogate dual problem. This method no longer requires alternative iterations over the primal and dual spaces, as the previous primal-dual algorithm does, and can therefore save much computational effort. An algorithmic procedure of this dual method is suggested.

6.1 INTRODUCTION

The structural optimization problem considered here consists of the weight minimization of a structure with fixed geometry and material properties. The transverse sizes of the structural members, such as bar cross-sections, plate thicknesses, etc., are the design variables denoted by a_i . They are subjected to the size constraints

$$a_i \geq a_i' \quad i=1, \dots, n \quad (6.1)$$

where a_i' are lower bounds assigned to the design variables and n is the number of independent design variables after linking. The behaviour constraints impose limitations on quantities describing the responses of the structure to the applied forces, such as stresses and displacements under multiple static loading cases, the natural frequencies, etc. They can be written as nonlinear inequalities:

$$u_j(a) \leq u_j' \quad j=1, \dots, m \quad (6.2)$$

where $u_j(a)$ denotes an implicit structural response function of the design variables that has to be lower than a given limit u_j' . The objective function to be minimized is the structural weight. It is a linear function of the design variables:

$$W = \sum_{i=1}^n \rho_i l_i a_i \quad (6.3)$$

where ρ_i is the mass density and ℓ_i is a geometrical parameter such that the product $\ell_i a_i$ is the volume of the i th linking group. The optimization problem thus consists in minimizing the objective function (6.3) subject to the constraints (6.1) and (6.2). It is a nonlinear mathematical programming problem.

The difficulty in solving the problem arises from the fact that the $u_j(a)$ appearing in Eq.(6.2) are in general implicit functions of the design variables that must be evaluated by executing a complete finite element analysis. This difficult implicit nonlinear programming problem is usually solved by transforming it into a sequence of simpler explicit problems. In these approximate problems, the real constraints are replaced with first-order explicit approximations

$$u_j(a) = \sum_{i=1}^n c_{ij}/a_i \quad j=1, \dots, m \quad (6.4)$$

where the coefficients c_{ij} are evaluated after each structural analysis. They are considered as constant in the optimization phase. Therefore, the explicit problem to be solved after each structural analysis has the following form:

(P0)

$$\text{Min} \quad W = \sum_{i=1}^n \rho_i \ell_i a_i \quad (6.3)$$

$$\text{s.t.} \quad \sum_{i=1}^n c_{ij}/a_i - u_j' \leq 0 \quad j=1, \dots, m \quad (6.5)$$

$$a_i' - a_i \leq 0 \quad i=1, \dots, n \quad (6.1)$$

It should be noted that for a statically determinate structure subject to static stress and displacement constraints only, problem (P0) is the exact explicit form of the "real" problem, since in this case c_{ij} remain constant coefficients. The minimum weight design is obtained in only one structural analysis as the solution of problem (P0). In the case of a statically indeterminate structure, the c_{ij} are implicit functions of the design variables and they can be known only numerically by means of a structural analysis. In this case, problem (P0) is solved after each structural analysis by considering c_{ij} as constant coefficients. The structure is then reanalysed and problem (P0) is updated and solved. This process is repeated until convergence is achieved.

The explicit constraints (6.5) are, in fact, linearized forms of the original constraints (6.2), resulting from a first-order Taylor series expansion w.r.t. the reciprocal design variables $x_i = 1/a_i$. Then, problem (P0) can be further replaced by problem (P) in which the problem variables are the reciprocal design variables x_i and the problem constraints are expressed in terms of the variables x_i and rearranged. Consequently, problem (P) takes the following form:

(P)

$$\text{Min } W = \sum_{i=1}^n \rho_i \ell_i / x_i \quad (6.6)$$

$$\text{s.t. } g_j = \sum_{i=1}^n c_{ij} x_i / u_j - 1 \leq 0 \quad j=1, \dots, m \quad (6.7)$$

$$g_{m+i} = a_i x_i - 1 \leq 0 \quad i=1, \dots, n \quad (6.8)$$

which is a convex programming problem with only linear constraints. There have been many methods proposed to solve either problem (P0) or problem (P). They range from mathematical programming techniques to optimality criteria approaches (Refs.[9], [40-43], [78]). It is now widely accepted that the special structure of this problem permits a dual formulation. The Lagrange multipliers associated with the constraints are the dual variables in terms of which an auxiliary and equivalent problem can be stated. In addition, the separable property of the problem functions leads the dual formulation to an efficient solution scheme since each primal variables can be independently expressed in terms of the dual variables. The derived dual problem can then be solved by the Newton-Raphson method, as suggested by Fleury[40]. However, discontinuity planes for the second derivatives of the dual function exist in the dual space such that some precautions must be taken in order to avoid the singularity of the Hessian matrix. Furthermore, the dual algorithms are usually subject to instability in the convergence of the structural weight. In the author's opinion, the main cause of these difficulties consists of the fact that almost all the algorithms, including some primal algorithms, have invariably adopted the so-called "active set strategies"; that is, during the course of the optimization process constraints which appear to be slack are temporarily deleted within an iteration to reduce the problem size. On exit the deleted constraints are checked and if they are violated

by the current point a new "active set" of constraints is chosen and a new iteration is performed. For a practical problem, therefore, it is possible for the active constraint set to vary during the optimization process. In the present work an active set strategy is not necessary and some optimization methods are developed in which none of the constraints is deleted during the optimization process. The computational effort required remains the same or increases only marginally.

In the next sections, several ways of solving the structural optimization problem under consideration are investigated. They employ the surrogate constraint approach and the entropy maximization or the equivalent entropy augmentation approach. As indicated in Chapter 4, whenever the surrogate problems can be effectively solved the primal-dual method developed on the basis of the entropy maximization approach can be used. Moreover, if an explicit surrogate dual is obtainable, the dual method based upon the idea of entropy augmentation can be applied. The structural optimization problem is such a case. The developments made in this chapter can, therefore, be viewed as applications of the methods presented in Chapter 4 to a particular class of optimization problems.

The organization of this chapter is as follows. In Section 6.2, a primal-dual method is proposed in which problem (P) is solved by means of solving a sequence of surrogate problems combined with an appropriate update scheme for the surrogate multipliers. Analytical solutions of the surrogate problems can be obtained using Lagrangean stationarity

conditions and expressed in terms of the surrogate multipliers, which are then updated by the entropy-based update formula developed in Chapter 4. Section 6.3 describes a primal-dual algorithm for solving the structural optimization problems. In Section 6.4, some example problems are presented and the computational results are compared with those previously reported by other authors (Refs.[43], [78], [85], [100]). Section 6.5 is devoted to deriving an explicit surrogate dual for the structural optimization problem by two ways, one of which uses the Lagrangean duality theory and the other makes use of geometric programming duality theory. Section 6.6 describes an entropy-based dual method to effectively solve the resulting dual problem, and hence the original problem (P). There is no doubt that the solution methods presented herein can compete with the best methods available for solving this class of structural optimization problems.

6.2 PRIMAL-DUAL METHOD FOR STRUCTURAL OPTIMIZATION

The method employs the concept of surrogate constraints. A surrogate problem (S) corresponding to problem (P) can be formulated as follows:

(S)

$$\text{Min } W = \sum_{i=1}^n \rho_i \ell_i / x_i \quad (6.6)$$

$$\text{s.t.} \quad \sum_{j=1}^m \lambda_j \left(\sum_{i=1}^n c_{ij} x_i / u_j - 1 \right) + \sum_{i=1}^n \lambda_{m+i} (a_i' x_i - 1) = 0 \quad (6.9)$$

where the surrogate multipliers satisfy the normality condition

$$\sum_{j=1}^m \lambda_j + \sum_{i=1}^n \lambda_{m+i} = 1 \quad (6.10)$$

The surrogate constraint has been written as an equality because in almost all structural optimization problems, there will be at least one constraint active at the optimum and consequently the surrogate constraint must be active as indicated by Proposition 1.3.4.

Introducing Lagrange multiplier α for the surrogate constraint (6.9), we have the surrogate Lagrangian of problem (S) as:

$$L(x, \lambda, \alpha) = \sum_{i=1}^n \rho_i \ell_i / x_i + \alpha \left[\sum_{j=1}^m \lambda_j \left(\sum_{i=1}^n c_{ij} x_i / u_j - 1 \right) + \sum_{i=1}^n \lambda_{m+i} (a_i' x_i - 1) \right] \quad (6.11)$$

The solution of problem (S) must satisfy the following Lagrangean stationarity conditions w.r.t. x_i

$$\partial L / \partial x_i = -\rho_i \ell_i / x_i^2 + \alpha \left[\sum_{j=1}^m c_{ij} \lambda_j / u_j + a_i' \lambda_{m+i} \right] = 0 \quad i=1, \dots, n \quad (6.12)$$

and the stationarity condition w.r.t. α , $\partial L / \partial \alpha = 0$, which leads to the surrogate constraint, Eq.(6.9). The analytical solution of problem (S)

can be readily found from (6.12) and expressed as functions of the surrogate multipliers:

$$x_i = \sqrt{\rho_i \ell_i / \alpha u_i} \quad i=1, \dots, n \quad (6.13)$$

where

$$u_i = \sum_{j=1}^m c_{ij} \lambda_j / u_j + a_i \lambda_{m+i} \quad i=1, \dots, n \quad (6.14)$$

To find the unknown α in Eq.(6.13), we substitute the x_i from (6.13) into (6.9) so that α can be obtained as:

$$\alpha = \left[\sum_{i=1}^n \sqrt{\rho_i \ell_i u_i} \right]^2 \quad (6.15)$$

which replaces the unknown α of (6.13) to yield the reciprocal design variables x_i as

$$x_i = \sqrt{\rho_i \ell_i / u_i} / \sum_{i=1}^n \sqrt{\rho_i \ell_i u_i} \quad i=1, \dots, n \quad (6.16)$$

and the primal design variables then are

$$a_i = \sqrt{u_i / \rho_i \ell_i} / \sum_{i=1}^n \sqrt{\rho_i \ell_i u_i} \quad i=1, \dots, n \quad (6.17)$$

Thus we have derived the redesign formula in terms of the surrogate multipliers and, therefore, reduced the structural optimization problem to one in which an efficient update scheme for the multipliers is needed.

To treat all the constraints consistently, as explained in Chapter 4, the updates of the multipliers should follow an entropic maximization process which, for the structural optimization problem under consideration, leads to the following update formula

$$\lambda_j = \exp(p\alpha g_j) / \sum_{j=1}^{m+n} \exp(p\alpha g_j) \quad j=1, \dots, m+n \quad (6.18)$$

where p is a control parameter which increases with the optimization process. The procedure for deriving (6.18) follows that illustrated in Chapter 4 and is not repeated here.

A primal-dual method can then be developed in which alternate iterations are carried out over both x -space and λ -space. The method starts with setting all the surrogate multipliers to be equal, i.e., $\lambda_j = 1/(m+n)$, which corresponds to setting the control parameter $p=0$. The analytical solution x_i of problem (S) are then evaluated by Eqs.(6.16). Values of α and the constraints ($j=1, \dots, m+n$) are also calculated by Eqs.(6.15), (6.7) and (6.8). The surrogate multipliers λ_j are then updated using Eq.(6.18) accompanied by an appropriate increase of the parameter p . Thus, the process is ready for another analytical solution

given by Eq.(6.16). This procedure is repeated until convergence is reached; that is, the solution of problem (P) is found.

6.3 PRIMAL-DUAL ALGORITHM FOR STRUCTURAL OPTIMIZATION

This section is devoted to describing an entropy-based primal-dual algorithm which is devised on the basis of the primal-dual method presented in the last section and is, in fact, an application of the primal-dual algorithm, as developed in Chapter 4, to a particular class of nonlinear programming problems: the structural optimization of pin-jointed trusses. The algorithm is divided into two phases, structural analysis and structural optimization. The latter is further subdivided into two stages, the solution of problem (S) (primal) and the updates of the surrogate multipliers (dual).

The primal-dual algorithm developed to solve the structural optimization problem in question is described as follows:

1. Input geometrical and structural data, and set the global iteration number $kk=0$.
2. Set $kk=kk+1$, and carry out a structural analysis by the matrix stiffness method to evaluate structural responses such as node displacements, member forces, etc.

3. Scale the current design: $a' = \gamma a$ where γ is a scaling coefficient to make the scaled design feasible and at least one of the constraints active. This ensures that a feasible design is always available whenever the design process terminates.
4. Check global convergence criteria and if they are satisfied the algorithm stops.
5. Enter the optimization phase in which there are alternate iterations over the two spaces of the design variables and the surrogate multipliers. The procedure within this phase is described below.
 - a. Set the local iteration number $k=0$ and the parameter $p=0$, i.e., $\lambda_j = 1/(m+n)$, ($j=1, \dots, m+n$).
 - b. Set $k=k+1$ and evaluate the design variables x_i ($i=1, \dots, n$) by Eq.(6.16); calculate the structural weight W^k and the constraint values using Eqs.(6.6), (6.7) and (6.8) as well as α by (6.15).
 - c. Check local convergence criteria and if they are satisfied the optimization phase (the local iteration) terminates.
 - d. Increase the control parameter p by a positive increment and update the surrogate multipliers λ_j by Eq.(6.18).
 - e. Go to Step b.
6. Assign the design variables $a_i^{kk} = 1/x_i^{kk}$ ($i=1, \dots, n$) which are obtained from the optimization phase, and go to Step 2.

Some explanations on the proposed algorithm are necessary. They are summarized as:

1. The optimization problem to be solved is formulated at the scaled design point where the stress constraints are combined with the size

constraints to give the modified size constraints a_i' , and the coefficients c_{ij} appearing in the displacement constraints are calculated by the virtual work principle. The a_i' and c_{ij} all remain constants in the optimization phase.

2. The global convergence criteria used in the example problems are
 - a. The algorithm terminates whenever the structural weight evaluated at the scaled design point is not less than that obtained in the last iteration, i.e.,

$$w^{kk} \geq w^{kk-1}$$

where w^{kk} and w^{kk-1} are the structural weights of the current iteration and the last iteration, respectively.

- b. The relative improvement of the structural weights at two subsequent scaled design points is less than a specified value $\epsilon_1=10^{-4}$ for instance, i.e.,

$$(w^{kk-1}-w^{kk})/w^{kk-1} < \epsilon_1$$

- c. The relative error of the design variables is less than a specified value $\epsilon_2=10^{-3}$, i.e.,

$$\|x^{kk}-x^{kk-1}\| \leq \epsilon_2$$

3. The local convergence criteria are

- a. The optimization phase terminates whenever the structural weight of the current iteration is not greater than that obtained in the last iteration, i.e.,

$$w^k \leq w^{k-1}$$

which is opposite to the global criterion (a) as the iterations within this phase correspond to the solution of a dual problem, i.e., the maximization over the dual variables (the surrogate multipliers).

- b. The relative error of the surrogate multipliers is less than a specified value $\varepsilon_2=10^{-3}$, i.e.,

$$\|\lambda^k - \lambda^{k-1}\| \leq \varepsilon_2$$

- c. The global criteria b) and c) are also used as the local criteria.
4. The control parameter p increases by a small amount $\delta p=0.01$ at a time.

6.4 NUMERICAL EXAMPLES

Three examples are selected such that the results can be compared with those previously reported (Refs.[43], [78], [85], [94], [101]). The

solution methods used there consist of optimality criterion methods and mathematical programming methods. The entropy-based primal-dual algorithm described in the last section is used to solve these example problems. Comparisons are made of results and iteration histories.

Example 6.1: The first example is a ten-bar truss as shown in Fig.6.1. Data required are also given in Fig.6.1. The truss is designed to satisfy stress, displacement and prespecified minimum size constraints. No variable linking has been used in this example. Results are given in Table 6.1 along with some comparisons from the literature.

The minimum weight obtained is 5075.1 lbs in 16 reanalyses. The iteration history is depicted in Fig.6.2 which shows that the present method has better convergence performance than the others, especially in the first few iterations.

Example 6.2: The second example is a 25-bar space truss, the transmission tower. The structure and the node as well as member numbering are illustrated in Fig.6.3 along with the design data. Table 6.2 includes design variable linking and compression stress limits. Design variable linking is used to impose symmetry with respect to both the y-z and the x-z planes. The structure is subjected to two distinct load cases as given in Table 6.3.

The results are summarized in Table 6.4 together with some comparisons with those previously reported. A minimum weight of 545.38 is obtained in only 5 iterations. The iteration history is shown in

Fig.6.4 which features a great decrease of the weight in the first couple of iterations.

Example 6.3: The third example is a 72-bar space truss. Fig.6.5 shows the geometry of the structure, and the node and member numbering system are illustrated in detail for the uppermost tier. The loading systems are given in Table 6.5. The symmetry of the structure and its loading conditions are such that the number of independent design variables can be reduced to 16 using variable linking. The material properties, allowable stresses and minimum member sizes are given in Fig.6.5. The displacements of nodes 1-4 are limited to ± 0.25 in. in both x and y directions.

The results are summarized in Table 6.6 along with comparisons with the other methods. A minimum weight of 379.666 lbs is obtained in 6 iterations. The iteration history is shown in Fig.6.6, and is, as with the first two examples, favourable to the present algorithm.

It has been shown, from the results and the comparisons made with the other algorithms for all the three example problems, that the entropy-based primal-dual algorithm is very efficient in solving this class of structural optimization problems. Especially in the first few iterations, the structural weights of feasible designs are reduced more quickly than those obtained by the other algorithms, as shown from the figures for the iteration history. The algorithm displays better stability than the usual dual algorithms. This is achieved through the consistent treatment of all the constraints in the optimization process.

Efficiency of the algorithm in the optimization phase is independent of both the number of design variables and the number of constraints. This makes the algorithm suitable for the optimum design of large-scale structures. The separable property of the problem functions, of course, is another reason for the success of the algorithm. For the case of general finite element optimization, it is unlikely that analytical solutions such as that obtained for the structural optimization problem under consideration could be obtained. Nevertheless, the updates of surrogate multipliers based upon entropic processes remain effective.

6.5 SURROGATE DUAL FOR STRUCTURAL OPTIMIZATION

The structural optimization problems considered in the previous sections have been formulated as convex separable programming problems for which the corresponding surrogate problems have analytical solutions in terms of the given surrogate multipliers, as shown in Section 6.2. It is then possible to derive an explicit surrogate dual problem. In this section two ways of deriving the dual are presented, one of which uses the Lagrangean stationarity conditions and the other makes use of GP duality theory. It turns out that the derived dual problem is exactly the same as that proposed by Templeman[91]. The solution of the derived dual problem is left to Section 6.6.

6.5.1 LAGRANGEAN DUAL FORMULATION

In section 6.2, we obtained the analytical solution of problem (S) using Lagrangean stationarity conditions. The remaining work for deriving the surrogate dual is trivial. All we have to do is to substitute the analytical solution into the surrogate Lagrangean (6.11). On substituting the x_i ($i=1, \dots, n$) and α from Eqs.(6.16) and (6.15) into Eq.(6.11), the surrogate dual $s(\lambda)$ is obtained as the following explicit form:

$$s(\lambda) = V^2(\lambda) = \left[\sum_{i=1}^n \sqrt{\rho_i \ell_i} v_i \right]^2 \quad (6.19)$$

where v_i are given by Eq.(6.14). Consequently, the surrogate dual problem (SD) consists in maximizing $s(\lambda)$ subject to a single normality constraint (6.10), i.e.,

(SD)

$$\text{Max} \quad s(\lambda) = \left[\sum_{i=1}^n \sqrt{\rho_i \ell_i} v_i \right]^2 \quad (6.19)$$

$$\text{s.t.} \quad \sum_{j=1}^{m+n} \lambda_j = 1 \quad (6.10)$$

Alternatively, one may prefer using the square root $V(\lambda)$ of $s(\lambda)$ as the dual objective function to be maximized. Thus an equivalent surrogate dual problem (SD') is posed as

(SD')

$$\text{Max } V(\lambda) = \sum_{i=1}^n \sqrt{p_i l_i v_i} \quad (6.20)$$

$$\text{s.t. } \sum_{j=1}^{m+n} \lambda_j = 1 \quad (6.10)$$

It is obvious that problems (SD) and (SD') have an identical solution so that they are used interchangeably later. It is not difficult to recognize that problem (SD') is exactly the same as that proposed by Templeman[91] under some notational changes. Before describing the solution of problem (SD), we present an alternative way to derive the explicit surrogate dual $s(\lambda)$.

6.5.2 GEOMETRIC PROGRAMMING DUAL FORMULATION

To derive the surrogate dual $s(\lambda)$ by means of the GP duality theory, we have to first rearrange the terms in the surrogate constraint (6.9) by collecting the coefficients of the variables x_i to yield an alternative form of the surrogate constraint. The objective function (6.6) together with the rearranged surrogate constraint poses an alternative form of the surrogate problem (S) as

(S')

$$\text{Min } W = \sum_{i=1}^n \rho_i \ell_i / x_i \quad (6.6)$$

$$\sum_{i=1}^n v_i x_i \leq 1 \quad (6.21)$$

where v_i are given by Eq.(6.14). Thus, the surrogate problem (S') consists in maximizing the objective function (6.6) subject to a single constraint (6.21) in which both the objective and constraint functions have the standard form of posynomials required by GP theory. One can then employ GP duality theory[26] to establish the duality relationships between the surrogate problem (S') and its GP dual problem (DG) which is written as

(DG)

$$\text{Max } d(\omega) = \prod_{i=1}^n (\rho_i \ell_i / \omega_{0i})^{\omega_{0i}} \prod_{i=1}^n (v_i \omega_{10} / \omega_{1i})^{\omega_{1i}} \quad (6.22)$$

$$\text{s.t. } \sum_{i=1}^n \omega_{0i} = 1 \quad (6.23)$$

$$\sum_{i=1}^n \omega_{1i} = \omega_{10} \quad (6.24)$$

$$-\omega_{0i} + \omega_{1i} = 0 \quad i=1, \dots, n \quad (6.25)$$

where ω_{0i} and ω_{1i} ($i=1, \dots, n$) are the dual variables introduced for each term in the objective and constraint function, respectively. From (6.25), it follows immediately that

$$\omega_{1i} = \omega_{0i} \quad i=1, \dots, n \quad (6.26)$$

Substituting ω_{1i} from (6.26) into (6.24) and noting (6.23) give

$$\omega_{10} = 1. \quad (6.27)$$

Substituting ω_{1i} and ω_{10} from (6.26) and (6.27) into (6.22) and collecting the terms give

$$d(\omega) = \prod_{i=1}^n (\psi_i / \omega_{0i})^{2\omega_{0i}} \quad (6.28)$$

where

$$\psi_i = \sqrt{\rho_i \ell_i v_i} \quad (6.29)$$

Taking the logarithm of $d(\omega)$ in Eq.(6.28) yields

$$\ln d(\omega) = \sum_{i=1}^n 2\omega_{0i} [\ln(\psi_i) - \ln(\omega_{0i})] \quad (6.30)$$

The GP dual problem defined by Eqs.(6.22)-(6.25) has reduced to a simpler problem in which the function (6.30) is maximized subject to the normality condition (6.23). The solution to this simplified GP dual problem can be obtained using its Lagrangean stationarity conditions as

$$\omega_{0i} = \psi_i / \sum_{i=1}^n \psi_i \quad i=1, \dots, n \quad (6.31)$$

which are then substituted into (6.28) to yield the surrogate dual as

$$s(\lambda) = \left[\sum_{i=1}^n \psi_i \right]^2 \quad (6.32)$$

which can be easily recognized to be the same as that given by Eq.(6.19) if considering Eq.(6.29).

In the above derivation, several points are of interest. Firstly, the equivalent surrogate problem (S') has taken a surprisingly simpler form than problem (S). Secondly, the GP dual problem defined by Eqs.(6.22)-(6.25) can be analytically solved by simple algebraic operations. The third is that the Lagrangean duality and the GP duality once again produce an identical result as done in Chapter 3.

6.6 ENTROPY-BASED DUAL METHOD FOR STRUCTURAL OPTIMIZATION

It should be noted that although the surrogate dual problem (SD) or (SD') appears to be simple, there is no effective algorithm for solving them. The only algorithm available to date was proposed by Sindaha[88]

in which only the displacement constraints were considered in the early stages of the optimization process. The algorithm has less rigorous background and consequently suffers from instability.

The developments in this chapter are towards providing not only alternative derivations of the dual problem but also its solution. As the derived surrogate dual function has an explicit form, the entropy-based dual method presented in Chapter 4 is preferred. In fact, the solution method to be described is merely an application of the entropy based dual method of Chapter 4 to a particular class of optimization problems, namely structural optimization problems. As indicated in Chapter 4, the entropy-based dual method will be very efficient if an explicit surrogate dual can be obtained. This is the case for the structural optimization problem (P).

In what follows, we give the formulas for developing entropy-based algorithms. A recurrence formula (4.31) for the surrogate multipliers λ_j has been derived and takes, for this structural optimization problem, the form:

$$\lambda_j = \exp[p \partial s(\lambda) / \partial \lambda_j] / \sum_{j=1}^{m+n} p \partial s(\lambda) / \partial \lambda_j \quad j=1, \dots, m+n \quad (6.33)$$

in which p is a control parameter as before and the partial derivatives $\partial s(\lambda) / \partial \lambda_j$ can be evaluated by

$$\partial s(\lambda)/\partial \lambda_j = V(\lambda) \sum_{i=1}^n (c_{ij}/u_j)' \sqrt{\rho_i \ell_i / v_i} \quad j=1, \dots, m \quad (6.34)$$

and

$$\partial s(\lambda)/\partial \lambda_j = V(\lambda) a_{j-m}' \sqrt{\rho_{j-m} \ell_{j-m} / v_{j-m}} \quad j=m+1, \dots, m+n \quad (6.35)$$

where $V(\lambda)$ and v_i are given by Eqs.(6.20) and (6.14), respectively.

Eq.(6.33) together with Eqs.(6.34) and (6.35) constitute a basis for developing numerical algorithms for solving the structural optimization problem (P). The solution to problem (SD) or (SD') includes two phases: a local iteration phase and a global iteration phase. The reason for requiring the local iterations lies in the fact that Eq.(6.33) is solely a recurrence formula rather than an analytical solution. Within the local iteration phase, the parameter p remains constant. An algorithmic procedure for solving the surrogate dual problem (SD) or (SD') is suggested as follows:

1. Specify the maximum local iteration number $N_{\max}=10$, for instance, and $\delta p > 0$.
2. Set the global iteration number $kk=1$ and $p=0$, which corresponds to setting $\lambda_j=1/(m+n)$.
3. Enter the local iteration phase by setting the local iteration number $k=1$. The procedure of this phase is described below.
 - a. Update λ_j using the recurrence formula (6.33).
 - b. Check if $k > N_{\max}$, if so, go to Step (e).

- c. Check the local convergence criteria, if they are satisfied, go to Step 4.
 - d. Set $k=k+1$ and return to Step (a) but with the updated λ_j .
 - e. Set $\delta p = \gamma \delta p$ ($\gamma < 1$) and reset $k=1$, and then go to Step (a).
4. Check the global convergence criteria, if they are satisfied, the algorithm stops.
 5. Set $kk=kk+1$ and $p=p+\delta p$, then go to Step 3.

The global convergence criteria are suggested as

1. The algorithm stops whenever $s^{kk} \leq s^{kk-1}$. This criterion has been implied by Eq.(4.18), a result of Theorem 4.1. Since $p^{kk} > p^{kk-1}$, it should follow that $s^{kk} > s^{kk-1}$ according to (4.18). The case that $s^{kk} \leq s^{kk-1}$ occurs only if the maximizer has been approached in the presence of numerical errors.
2. The relative improvement $(s^{kk} - s^{kk-1})/s^{kk-1}$ is less than a specified $\epsilon_1 = 10^{-4}$.
3. The norm $\|\lambda^{kk} - \lambda^{kk-1}\|$ is less than a specified $\epsilon_2 = 10^{-3}$.

The global convergence criteria (2) and (3) are used as the local convergence criteria, but different values of ϵ_1 and ϵ_2 may be specified.

Generally speaking, a small δp requires fewer local iterations and makes the convergence stable. A large δp requires fewer global iterations, but the requirement for reducing δp may take place.

However, as proved by Theorem 4.1, an appropriate sequence of p ensures the convergence of the algorithm.

6.7 CONCLUSIONS

In the previous sections, the structural optimization problem (P) has been solved by the entropy-based primal-dual algorithm proposed in this chapter, which has demonstrated fast and stable convergence as a result of the consistent treatment of the problem constraints. It can then be strongly argued that active set strategies guarantee neither stable nor fast convergence. Moreover, the present method does not require sophisticated mathematical operations so that it is very easy to program and both the number of variables and the number of constraints increase computational efforts only marginally. Thus, the present method is very suitable for the solution of optimum design problems of large-scale structures.

An explicit dual problem for the structural optimization problem under consideration was then derived in terms of the surrogate constraint approach and first recognized as a surrogate dual problem. This has justified Templeman's dual formulation[91] and cast new insights into his work. The derivations presented here are simpler and more rigorous.

The explicit dual problem (SD'), as proposed by Templeman, has not been fully appreciated mainly because there is no efficient algorithm for solving it. The entropy-based dual method presented in Section 6.6 has thus bridged the gap between the dual problem and its solution. This may attract more attention to the dual formulation.

The structural optimization problem (P) has been investigated by many researchers and a great deal of solution methods have been proposed which are roughly divided into two categories, the mathematical programming approach and the optimality criteria approach. A long and bitter debate between advocates of the two approaches stopped only recently[40]. The present development has greatly simplified the problem presentations and the solution methods. There is no doubt that the solution methods presented herein are among the best available solution methods at the time of writing.

It should be noted that the entropy-based dual method is more efficient than the primal-dual method as it does not require alternative iterations over both primal and dual spaces. It is unfortunate that this method has been a very late development in the present research so that the example problems have not been calculated by the corresponding algorithm. There is, however, no doubt about the efficiency and stable convergence of this method, as indicated by Theorem 4.1.

Table 6.1 Results for Ten-Bar Truss

No. of member	Final design(in ²)				
	Schmit[85]	Venkayya[101]	Gellatly[43]	Qian[78]	The present
1	30.670	30.416	31.350	30.902	30.930
2	0.100	0.128	0.100	0.100	0.100
3	23.760	23.408	20.030	23.545	23.597
4	14.590	14.904	15.600	14.960	14.964
5	0.100	0.101	0.140	0.100	0.100
6	0.100	0.101	0.240	0.297	0.287
7	8.578	8.696	8.350	7.611	7.659
8	21.070	21.084	22.210	21.275	21.282
9	20.960	21.077	22.060	21.156	21.162
10	0.100	0.186	0.100	0.100	0.100
Weight	5076.85	5084.9	5112.	5069.4	5075.1(lbs)
Iters.	13	26	19	12	16

Table 6.2 Compression Stress Limits

Group	Members	Stress limits(pci)
1	1	-35092
2	2-5	-11590
3	6-9	-17305
4	10,11	-35092
5	12,13	-35092
6	14-17	-6759
7	18-21	-6959
8	22-25	-11082

Table 6.3 Loading Systems(kps)

Load		Direction		
cases	Node	X	Y	Z
1	1	10	-5	1
	2	10	-5	0
	3	0	0	0.5
	6	0	0	0.5
2	1	20	-5	0
	2	-20	-5	0

Table 6.4 Results for Twenty-Five Bar Truss

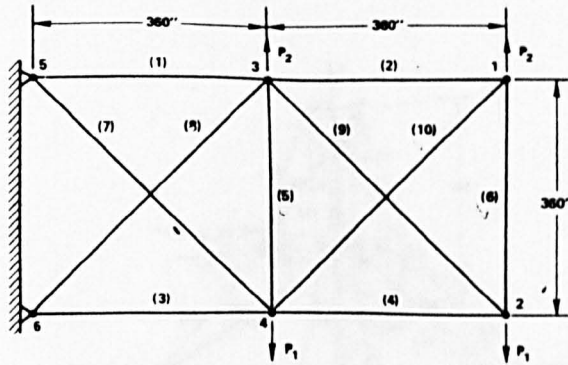
Group	Members	Schmit[85]	Final designs(in ²)		
			Templeman[94]	Gellatly[43]	The Present
1	1	0.010	0.010	0.010	0.010
2	2-5	1.964	2.022	2.007	1.976
3	6-9	3.033	2.934	2.963	2.994
4	10,11	0.010	0.010	0.010	0.010
5	12,13	0.010	0.010	0.010	0.010
6	14-17	0.670	0.670	0.688	0.684
7	18-21	1.680	1.675	1.678	1.687
8	22-25	2.670	2.697	2.664	2.663
Weight(lbs)		545.23	545.32	545.36	545.38
Iterations		16	7	7	5

Table 6.5 Applied Loading Systems to 72-Bar Truss(lbs)

Load cases	Node	Direction		
		X	Y	Z
1	1	5000	5000	-5000
2	1	0	0	-5000
	2	0	0	-5000
	3	0	0	-5000
	4	0	0	-5000

Table 6.6 Results for Seventy-Two Bar Space Truss

Group	Final designs(in ²)				
	Schmit[85]	Venkayya[101]	Gellatly[43]	Qian[78]	The Present
1	0.1565	0.161	0.1492	0.1564	0.1571
2	0.5458	0.557	0.7733	0.5457	0.5356
3	0.4105	0.377	0.4534	0.4106	0.4099
4	0.5699	0.506	0.3417	0.5692	0.5688
5	0.5233	0.611	0.5521	0.5237	0.5067
6	0.5173	0.532	0.6084	0.5171	0.5200
7	0.1000	0.100	0.1000	0.1000	0.1000
8	0.1000	0.100	0.1000	0.1001	0.1004
9	1.267	1.246	1.0235	1.2683	1.2801
10	0.5118	0.524	0.5421	0.5116	0.5148
11	0.1000	0.100	0.1000	0.1000	0.1000
12	0.1000	0.100	0.1000	0.1000	0.1000
13	1.8850	1.818	1.4636	1.8862	1.8973
14	0.5125	0.524	0.5207	0.5123	0.5158
15	0.1000	0.100	0.1000	0.1000	0.1000
16	0.1000	0.100	0.1000	0.1000	0.1000
Weight	379.640	381.2	395.97	379.62	379.666
Iters.	9	12	9	8	6



MATERIAL: ALUMINUM, $E = 10^7$ psi, $\rho = .1$ pci
 STRESS LIMITS: ± 25000 psi (ALL MEMBERS)
 LOWER LIMITS: .1 in² (ALL MEMBERS)
 UPPER LIMITS: NONE
 LOADING CASE 1: SINGLE LOAD $P_1 = 100$ K, $P_2 = 0$.
 LOADING CASE 2: SINGLE LOAD $P_1 = 150$ K, $P_2 = 50$ K

Fig.6.1* Ten-Bar Planar Truss

* Figs.6.1, 6.3 and 6.5 are quoted from Ref.[85].

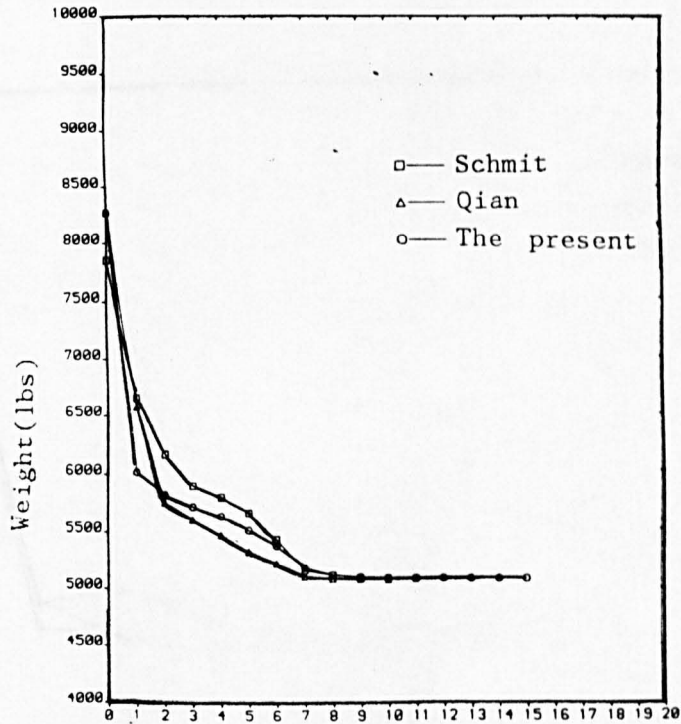
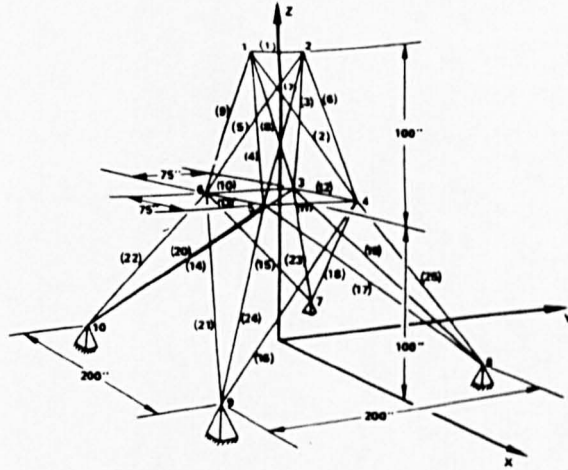


Fig.6.2 Iteration History of 10-Bar Truss



MATERIAL ALUMINUM, $E = 10^7$ psi, $\rho = .1$ psi
 STRESS LIMITS TENSION $+40000$ psi (SEE TABLE FOR COMPRESSION)
 LOWER LIMITS $.01$ in²
 UPPER LIMITS NONE
 DISPLACEMENT LIMITS $.35$ in. ON ALL NODES AND IN ALL DIRECTIONS

Fig.6.3 Twenty-Five-Bar Space Truss

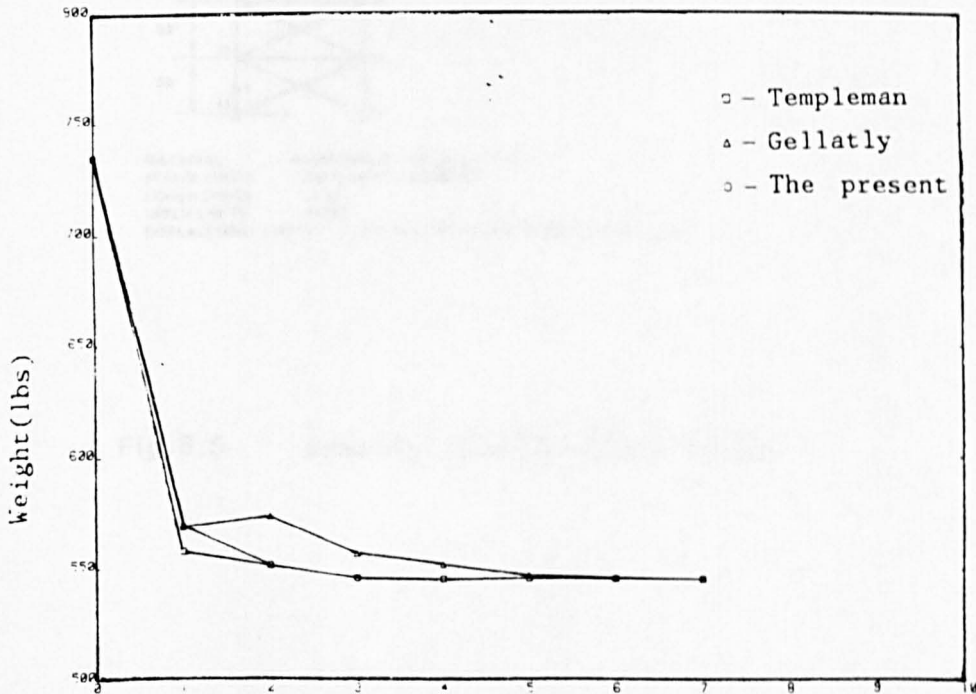


Fig.6.4 Iteration History of 25-Bar Truss

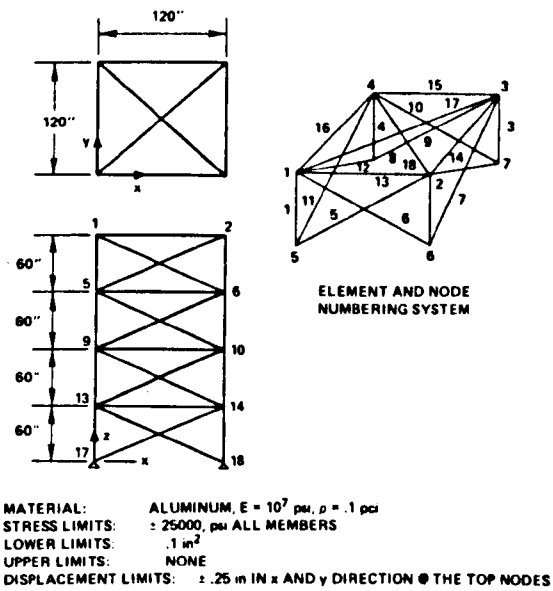


Fig.6.5 Seventy-Two Bar Space Truss

7.1 INTRODUCTION

This thesis analyzes the... measure and... solution of... two... the... optimization... domain... not... spontaneous... "temperature"... attacking...

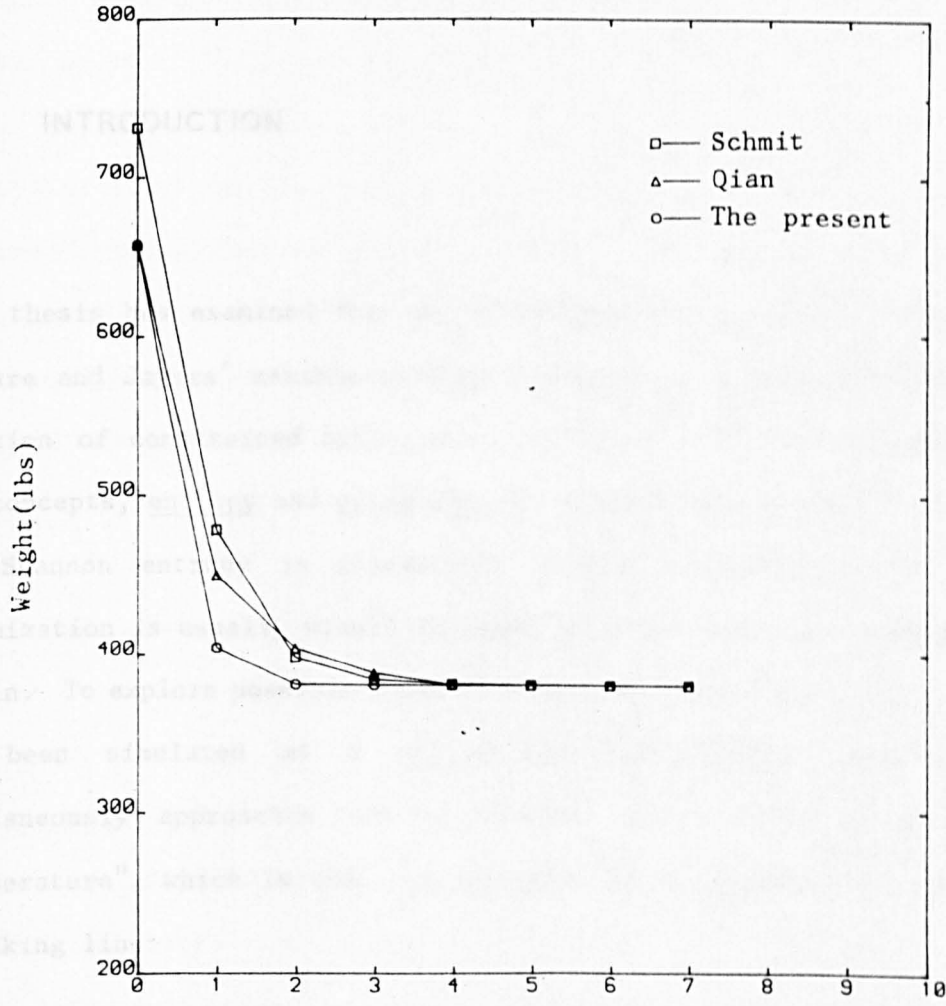


Fig.6.6 Iteration History of 72-Bar Truss

was then postulated... this thesis, the... "energy" of its... represented its total... of each wire state. The...

CHAPTER 7 CONCLUSIONS AND RECOMMENDATIONS FOR FUTURE WORK

7.1 INTRODUCTION

This thesis has examined the use of Shannon's (informational) entropy measure and Jaynes' maximum entropy criterion in connection with the solution of constrained optimization problems. At first glance, the two concepts, entropy and optimization, seem to have no direct link as the Shannon entropy is essentially related to probabilities while optimization is usually viewed in terms of a deterministic topological domain. To explore possible links between them, an optimization problem has been simulated as a statistical thermodynamic system that spontaneously approaches its equilibrium state under a specified "temperature", which is then characterized by the maximum entropy. An attacking line:

entropy \rightarrow thermodynamic equilibrium \rightarrow optimization

was then postulated. In the simulated thermodynamic system proposed in this thesis, the Lagrangean L_j in Eq.(4.8) was interpreted as the "energy" of its j th micro-state while the surrogate Lagrangean (4.7) represented its total energy which was the mean value of the energies of each micro-state. The surrogate multipliers thus have a probabilistic

interpretation. An optimum point then corresponds to a thermodynamic equilibrium state under a specified temperature. The entropy of the system as a measure of uncertainty contained in this process should attain a maximum value at the equilibrium state, and hence at the optimum point.

In the course of this study, the concept of entropy was further examined by presenting two dual forms. Several ways of using the entropy in an optimization context were investigated and their relations to some well-established results of optimization theory were studied. Some entropy-based optimization methods have been developed. A class of structural optimization problems was chosen to provide example problems.

The main developments made in the present study are summarized as follows:

1. Two dual forms of Jaynes' maximum entropy formalism were developed in terms of both Lagrangean duality theory and geometric programming duality theory. These dual formulations have provided additional insights into entropic processes as well as affording a simple means of calculating the least biased probabilities.
2. An entropy maximization model was proposed to estimate the probabilities of the simulated thermodynamic system which were represented by surrogate multipliers. It is then possible to provide optimization with an information-theoretic basis.
3. The idea of entropy augmentation was proposed, which justified the entropy maximization approach as well as the thermodynamic

simulation and provided a basis for developing a dual iterative method.

4. The entropy augmentation approach was further developed to yield a primal function for general constrained optimization problems, which can be used to develop new optimization algorithms.
5. Some properties of this primal function were studied by means of the arithmetic-geometric mean inequality and the concept of the norm of a vector. The differentiability and monotonically decreasing properties have suggested that good optimization algorithms could be developed using this function.
6. An entropy-based primal-dual algorithm for solving a class of structural optimization problems was developed and three numerical examples were presented.
7. An explicit surrogate dual for the above structural optimization problems was derived by means of both Lagrangean duality and geometric programming duality. An algorithmic procedure for a dual solution method was suggested.

7.2 CONCLUSIONS

The present work has shown that there are links between entropy and optimization. There is no doubt that good entropy-based optimization algorithms can be devised based upon the present development. Several conclusions, drawn from the present research, are summarized as follows:

1. The existence of the dual forms of the maximum entropy problem has provided not only an alternative convenient solution strategy but also additional insights into entropic processes. The recognition of a formal duality relationship is very important in using the maximum entropy criterion in nonequilibrium processes.
2. Uncertainty contained in the solution of constrained optimization problems is so similar to that contained in thermodynamic systems that it is reasonable to employ a statistical thermodynamic approach, i.e., the entropy maximization approach, to estimate surrogate multipliers.
3. The entropy augmentation approach, which may also be referred to as the entropy penalty approach, has a rigorous theoretic background, as proved by Theorem 4.1, and therefore promises stable convergence of the corresponding algorithms. Whenever an explicit surrogate dual exists the dual iterative method presented in Chapter 4 is very efficient.
4. The derivation of the primal function has a significance far beyond the function itself. A fact which must be emphasized is that it is the entropy augmentation approach which has made this possible. In the process of deriving this primal function, a half of the saddle-point problem has been analytically solved. This has greatly simplified the original problem, and has an effect similar to that of deriving an explicit dual. In other words, the minimization process has been realized if an explicit dual is obtained while the maximization process has been realized if an explicit primal is derived. It should be noted, however, that whether or not an explicit dual can be obtained depends upon the problem functions.

The answer is negative for most cases. The explicit primal function no longer has the above dependence. It therefore has wider applicability. As it has demonstrated the integrity and interactivity of the constraints it may have better performance than other penalty functions. Moreover, its differentiability and monotonically decreasing properties makes it easy to incorporate it in optimization algorithms.

5. The derivations of an explicit surrogate dual and the developments of the entropy-based methods have enabled the structural optimization problems considered in this thesis to be solved very easily. The results from three examples have shown that the primal-dual algorithm has fast and stable convergence, but better algorithms can be expected which iterate over dual space only. Through the developments made here, it can be seen that the surrogate dual formulation deserves to be more widely recognized than hitherto.

7.3 RECOMMENDATIONS FOR FUTURE WORK

The present work is exploratory. However it has opened up new avenues in the study of some classes of optimization problems, such as general minimax optimization and multicriteria optimization problems. Some potential research topics, which become possible due to the present work, are summarized as follows:

1. The present research is mainly oriented towards providing a theoretical basis for using the Shannon entropy in an optimization context. It has left many aspects of practical algorithmic development to be explored, which is equally important to the present work and are certainly required.
2. Further theoretical refinements are also required for more deeply understanding the present work and extending its applications to more optimization areas. Further links between entropy and optimization remain to be explored. It is now clear that the entropy-based approaches have close relations to minimax optimization. Thus the entropy augmentation approach can be readily adapted to solving minimax optimization problems by minimizing the minimax primal function F_p , as discussed in Section 5.5. This approach may also be used to solve other problems which are essentially of the minimax type, such as multicriteria optimization problems, curve fitting problems, equations and inequalities solving, etc. These remain to be explored.
3. The present work may also be extended to solve LP problems. It is known that the "surrogate cuts" approach can make the remaining ellipsoid small in the ellipsoid algorithm originated by Khachiyan[65]. However, the surrogate multipliers are generally found by solving a quadratic programming problem. The entropy-based updates should therefore find an immediate application to this problem with potentially large savings in computational effort. In Karmarkar's algorithm[64], an improvement can be made by adding an entropy term to the objective function. The additional entropy

process helps take into account the interactivity between the variables.

It may be possible for the primal function proposed in this thesis to be used to transform an LP problem into a sequence of pth problems which would then be solved by one of the three approaches proposed in Chapter 5.

4. The dual of the maximum entropy problem may be viewed as a generalized potential function. It may then be conjectured that a minimum potential problem may be replaced by a maximum entropy problem which is, in some sense and circumstance, easier to solve than the former. The key to realizing this objective is, however, how to give the problem in question an appropriate probabilistic interpretation. Some clues may be found in the present work.
5. Due to their minimax nature, the entropy-based approaches may be used to develop finite element methods, such as the least squares and penalty finite elements, along the line of the weighted residuals methods.

In conclusion, the main contribution to knowledge contained in this thesis centres around the various demonstrations that informational entropy and optimization processes are closely linked. Through this work it is now possible to view the traditional deterministic, topological interpretation of optimization processes as not the only interpretation, but as just one possible interpretation. It is equally admissible to treat optimization processes in a probabilistic, information-theoretic

way and to develop solution methods from this interpretation. This new insight opens up new avenues for research into optimization methods.

REFERENCES

1. N.Agmon, Y.Alhassid and R.D.Levine (1979), "An Algorithm for Determining the Lagrange Parameters in the Maximum Entropy Formalism," in **The Maximum Entropy Formalism** (edited by R.D.Levine and M.Tribus), MIT Press, 207-209.
2. J.S.Arora, E.J.Haug and S.D.Rajan (1981), "Efficient Treatment of Constraints in Large-Scale Structural Optimization," Eng. Opt., **5**, 105-120.
3. B.F.Backman (1976), "The Principle of Maximum Entropy with Applications to Problems in Structural Mechanics," Ph.D. Thesis, University of Washington, USA.
4. J.W.Bandler and C.Charalambous (1974), "Nonlinear Programming using Minimax Techniques," J. Opt. Theory Appl., **13**, 607-619.
5. P.C.Basu and A.B.Templeman (1984), "An Efficient Algorithm to Generate Maximum Entropy Distributions," Int. J. Numer. Methods in Eng., **20**, 1039-1055.
6. C.S.Beightler and D.T.Phillips (1976), **Applied Geometric Programming**, John Wiley.
7. M.Bellmore, H.J.Greenberg and J.J.Jarvis (1970), "Generalized Penalty Function Concepts in Mathematical Optimization," Opns. Res., **18**, 229-252.
8. A.Ben-Tal (1985), "The Entropic Penalty Approach to Stochastic Programming," Math. Opns. Res., **12**(2), 263-279.

9. L.Berke and N.S.Khot (1974), "Use of Optimality Criteria Methods for Large Scale Systems," AGARD-LS-70.
10. D.P.Bertsekas (1975), "Necessary and Sufficient Conditions for A Penalty Method to Be Exact," Math. Prog., **9**, 87-99.
11. D.P.Bertsekas (1975), "Combined Primal-Dual and Penalty Methods for constrained Minimization," SIAM J. Control, **13**, 521-544.
12. D.P.Bertsekas (1976), "Multiplier Methods: A Survey," Automatica, **12**, 133-145.
13. D.P.Bertsekas (1982), **Constrained Optimization and Lagrange Multiplier Methods**, Academic Press.
14. Bland R.G., Goldfarb D and Todd M.J.(1981), "The Ellipsoid Method: A Survey," Opns. Res., **29**, 1039-1091.
15. R.Brooks and A.Geoffrion (1966), "Finding Everett's Lagrange Multipliers by Linear Programming," Opns. Res., **14**, 1149-1153.
16. C.G.Broyden (1967), "Quasi-Newton Methods and Their Application to Function Minimization," Math. Comp., **21**, 368-381.
17. H.B.Callen (1960), **Thermodynamics**, John Wiley & Sons, New York.
18. V.Cerny (1985), "Thermodynamical Approach to the Traveling Salesman Problem: An Efficient Simulation Algorithm," J. Opt. Theory Appl., **45(1)**, 41-51.
19. C.Charalambous and J.W.Bandler (1976), "Nonlinear Minimax Optimization as A Sequence of Least pth Optimization with Finite Values of p," Int. J. Syst. Sci., **7**, 377-391.
20. C.Charalambous (1977), "Nonlinear least pth optimization and Nonlinear programming," Math. Prog., **12**, 195-225.

21. C.Charalambous (1979), "Acceleration of the Least pth Algorithm for Minimax Optimization with Engineering Applications," Math. Prog., **17**, 270-297.
22. V.F.Dem'yanov and V.N.Malozemov (1974), **Introduction to Minimax**, Ketev Publishing House, Jerusalem.
23. J.J.Dinkel, G.A.Kochenberger and S.-N.Wong (1977), "Entropy Maximization and Geometric Programming," Envir. and Planning A, **9**, 419-427.
24. J.J.Dinkel and G.A.Kochenberger (1978), "An Implementation of Surrogate Constraint Duality," Opns. Res., **26**, 358-364.
25. J.J.Dinkel and G.A.Kochenberger (1979), "Constrained Entropy Models: Solvability and Sensitivity," Manag. Sci., **25(6)**, 555-564.
26. R.J.Duffin, E.L.Peterson and G.Zener (1967), **Geometric Programming**, Wiley, New York.
27. J.Eriksson (1980), "A Note on Solution of Large Sparse Maximum Entropy Problems with Linear Equality Constraints," Math. Prog., **18**, 146-154.
28. S.Erlander (1977), "Accessibility, Entropy and the Distribution and Assignment of Traffic," Trans. Res., **11**, 149-153.
29. S.Erlander (1981), "Entropy in Linear Programs," Math. Prog., **21**, 137-151.
30. S.P.Evans (1973), "A Relationship between the Gravity Model for Trip Distribution and the Transportation Problem in Linear Programming," Trans. Res., **7**, 39-61.
31. H.Everett (III) (1963), "Generalized Lagrange Multiplier Method for Solving Problems of Optimum Allocation of Resources," Opns. Res., **11**, 399-471.

32. J.E.Falk (1967), "Lagrange Multipliers and Nonlinear Programming," J. Math. Anal. Appl., **19**, 141-159.
33. J.D.Fast (1968), **Entropy**, Philips Technical Library.
34. A.V.Fiacco and G.P.McCormick (1968), **Nonlinear Programming: Sequential Unconstrained Minimization Techniques**, John Wiley & Sons, New York.
35. R.Fletcher and C.M.Reeves (1964), "Function Minimization by Conjugate Gradients," Computer J., **7**, 149-154.
36. R.Fletcher (1970), "A New Approach to Variable Metric Algorithms," Computer J., **13**, 317-322.
37. R.Fletcher (1973), "An Exact Penalty Function for Nonlinear Programming with Inequalities," Math. Prog., **5**, 129-150.
38. R.Fletcher (1975), "An Ideal Penalty Function for Constrained Optimization," J. Inst. Math. Appl., **15**, 319-342.
39. R.Fletcher (1981), **Practical Methods of Optimization**, John Wiley & Sons.
40. C.Fleury (1979), "Structural Weight Optimization by Dual Methods of Convex Programming," Int. J. Numer. Methods in Eng., **14**, 1761-1783.
41. C.Fleury and L.A.Schmit (1980), "Primal and Dual Methods in Structural Optimization," ASCE J. Struct. Div., **106**(5), 1117-1133.
42. M.B.Fuchs and M.A.Brull (1978), "Norm Optimization Method in Structural Design," AIAA J., **16**(1), 20-28.
43. R.A.Gellatly and L.Berke (1970), "Optimum Structural Design," AFFDL-TR-70-165.
44. Geoffrion A.M.(1969), "An Improved Implicit Enumeration Approach for Integer Programming," Opns. Res., **17**, 437-454.

45. P.E.Gill and W.Murray (1974), "Newton-Type Methods for Unconstrained and Linearly Constrained Optimization," Meth. Prog., **28**, 311-350.
46. P.E.Gill and W.Murray(eds) (1974), **Numerical Methods for Constrained Optimization**, Academic Press, London.
47. P.E.Gill, W.Murray, M.A.Saunders, J.A.Tomlin and M.H.Wright (1986), "On Projected Newton Barrier Methods for Linear Programming and An Equivalence to Karmarkar's Projective Method," Math. Prog., **36(2)**, 183-209.
48. F.Glover (1965), "A Multiphase-Dual Algorithm for The Zero-One Integer Programming Problem," Opns. Res., **13**, 879-919.
49. F.Glover (1968), "Surrogate Constraints," Opns. Res., **16**, 741-749.
50. F.Glover (1975), "Surrogate Constraint Duality in Mathematical Programming," Opns. Res., **23**, 434-451.
51. F.J.Gould (1969), "Extensions of Lagrange Multipliers in Nonlinear Programming," SIAM J. Appl. Math., **13(6)**, 1280-1297.
52. H.J.Greenberg and W.P.Pierskalla (1970), "Surrogate Mathematical Programming," Opns. Res., **18**, 924-939.
53. D.S.Griffeath (1972), "Computer Solution of the Discrete Maximum Entropy Problem," Technometrics, **14(4)**, 891-897.
54. S.Guiasu (1977), **Information Theory with Applications**, McGraw-Hill, Inc.
55. S.Guiasu (1986), "Maximum Entropy Condition in Queueing Theory," J. Opl. Res. Soc., **37(3)**, 293-301.
56. P.Hajela (1982), "Techniques in Optimum Structural Synthesis with Static and Dynamic Constraints," Ph.D. Thesis, Stanford University, USA.

57. G.H.Hardy, J.E.Littlewood and G.Polya (1934), **Inequalities**, Cambridge University Press.
58. J.K.Hartman (1975), "Iterative Determination of Parameters for an Exact Penalty Function," J. Opt. Theory Appl., 16(1/2), 49-66.
59. M.R.Hestenes (1969), "Multiplier and Gradient Methods," J. Opt. Theory Appl., 4, 303-320.
60. E.T.Jaynes (1957), "Information Theory and Statistical Mechanics," Phys. Rev., 106, 620-630 and 108, 171-190.
61. E.T.Jaynes (1968), "Prior Probability," IEEE Trans. Sys., Sci. & Cyber., SSC-4(3), 227-241.
62. E.T.Jaynes (1979), "Where Do We Stand on Maximum Entropy?" in **The Maximum Entropy Formalism** (edited by Levine R.D. and Tribus M.), MIT Press, 15-118.
63. E.T.Jaynes (1983), **Papers on Probability, Statistics and Statistical Physics** (edited by R.D.Rosenkrantz).
64. N.Karmarkar (1984), "A New Polynomial-Time Algorithm for Linear Programming," Combinatorica, 4(4), 373-395.
65. L.G.Khachiyan (1979), "A Polynomial Algorithm in Linear Programming," Sov. Math. Doklady, 20(1), 191-194.
66. G.Kreisselmeier and R.Steinhauser (1979), "Systematic Control Design by Optimizing A Vector Performance Index," IFAC Symp. Computer Aided Design of Control Systems, Zurich, Switzerland.
67. H.W.Kuhn and A.W.Tucker (1951), "Nonlinear Programming," in Proc. of the 2nd Berkeley Symp. in Mathematical Statistics and Probability, (edited by J.Neyman).
68. L.S.Lasdon (1970), **Optimization Theory for Large Systems**, Macmillan, N.Y.

69. R.D.Levine and M.Tribus (eds) (1979), **The Maximum Entropy Formalism**, MIT Press.
70. D.G.Luenberger (1968), "Quasi-Convex Programming," SIAM J. Appl. Math., **16**, 1090-1095.
71. D.G.Luenberger (1973), **Introduction to Linear and Nonlinear Programming**, Addison-Wesley Publishing Company.
72. J.Munro and P.W.Jowitt (1978), "Decision Analysis in the Ready-Mixed Concrete Industry," Proc. Inst. Civ. Engs. Part 2, **65**, 41-52.
73. J.Munro (1984), "Fuzzy Programming and Imprecise Data," Civ. Eng. Sys., **1(5)**, 255-260.
74. A.Osyczka (1984), **Multicriterion Optimization in Engineering**, Ellis Horwood, Chichester.
75. J.G.Pierce (1979), "A New Look at the Relation between Information Theory and Search Theory," in **The Maximum Entropy Formalism** (edited by R.D.Levine and M.Tribus), MIT Press, 339-402.
76. T.Pietrzykowski (1969), "An Exact Potential Method for Constrained Maxima," SIAM J. Numer. Anal., **6**, 217-238.
77. M.J.D.Powell (1969), "A Method for Nonlinear Constraints in Minimization Problems," in **Optimization** (edited by R.Fletcher), 283-298, Academic Press, N.Y.
78. L.X.Qian, W.X.Zhong, Y.K.Sui and J.D.Zhang (1982), "Efficient Optimum Design of Structures - Program DDDU," J. Comp. Methods in Appl. Mech. and Eng., **30**, 209-224.
79. R.T.Rockafellar (1973), "The Multiplier Method of Hestenes and Powell Applied to Convex Programming," J. Opt. Theory Appl., **12**, 555-562.

80. R.T.Rockafellar (1973), "A Dual Approach to Solving Nonlinear Programming Problems by Unconstrained Optimization," Math. Prog., **5**, 354-373.
81. R.T.Rockafellar (1974), "Augmented Lagrange Multiplier Functions and Duality in Nonconvex Programming," SIAM J. Control, **12**, 268-285.
82. J.B.Rosen (1960), "The Gradient Projection Method for Nonlinear Programming, Part I: Linear Constraints," SIAM J., **8**, 181-217.
83. J.B.Rosen (1961), "The Gradient Projection Method for Nonlinear Programming, Part II: Nonlinear Constraints," SIAM J., **9**, 514-532.
84. G.Sander and C.Fleury (1978), "A Mixed Method in Structural Optimization," Int. J. Numer. Methods in Eng., **13**, 385-404.
85. L.A.Schmit and B.Farshi (1974), "Some Approximation Concepts for Structural Synthesis," AIAA J., **12**(5), 692-699.
86. C.E.Shannon (1948), "The Mathematical Theory of Communication," Bell Sys. Tech. J., **27**, 279-428 and 623-656.
87. J.N.Siddall and Y.Diab (1974), "The Use in Probabilistic Design of Probability Curves Generated by Maximizing the Shannon Entropy Function Constrained by Moments," ASME Paper No.74-WA/DE-2.
88. S.Y.Sindaha (1980), "Optimization of Trusses using A Dual Method," M.Eng. Thesis, The University of Liverpool, U.K.
89. J.Sobieszczanski-Sobieski (1986), "Multilevel Structural Optimization," NATO/NASA/NSF/USAF ASI on Computer Aided Optimal Design: Structural and Mechanical Systems, Troia, Portugal, Vol.3, 7-28.
90. J.Stoer (1963), "Duality in Nonlinear Programming and the Minimax Theorem," Numerische Mathematik, **5**, 371-379.

91. A.B.Templeman (1976), "A Dual Approach to Optimum Truss Design," J. Struc. Mech., 4(3), 235-255.
92. A.B.Templeman and X.S.Li (1985), "Entropy Duals," Eng. Opt., 9(2), 107-119.
93. A.B.Templeman and X.S.Li (1987), "A Maximum Entropy Approach to Constrained Nonlinear Programming," Eng. Opt., 12, 191-205.
94. A.B.Templeman and S.K.Winterbottom (1973), "Structural Design Applications of Geometric Programming," AGARD Conference Proceedings No.123, Second Symposium on Structural Optimization, Milan, Italy.
95. C.L.Tien and J.H.Lienhard (1979), **Statistical Thermodynamics**, McGraw-Hill.
96. J.F.Traub (1964), **Iterative Methods for The Solution of Equations**, Prentice-Hall.
97. M.Tribus (1961), "Information Theory as the Basis for Thermostatistics and Thermodynamics," ASME J. Appl. Mech., Paper No.60-WA-23, 1-8.
98. M.Tribus (1961), **Thermostatistics and Thermodynamics**, D. Van Nostrand Company Inc.
99. M.Tribus (1969), **Rational Descriptions, Decisions and Designs**, Pergamon, New York.
100. M.Tribus (1979), "Thirty Years of Information Theory," in **The Maximum Entropy Formalism** (edited by R.D.Levine and M.Tribus), MIT Press, 1-14.
101. V.B.Venkayya, N.S.Khot and L.Berke (1973), "Application of Optimization Criteria Approaches to Automated Design of Large Practical Structures," AGARD-CP-123.

102. R.Wait (1979), **The Numerical Solution of Algebraic Equations**, John Wiley & Sons.
103. P.Whittle (1971), **Optimization under Constraints**, John Wiley & Sons.
104. A.G.Wilson and M.L.Senior (1974), "Some Relationships between Entropy Maximizing Models, Mathematical Programming Models, and Their Duals," J. Regl. Sci., 14(2), 207-215.
105. P.Wolfe (1961), "A Duality Theorem for Non-linear Programming," Quart. Appl. Math., 19, 239-244.
106. W.I.Zangwill (1967), "Nonlinear Programming via Penalty Functions," Manag. Sci., 13(5), 344-358.
107. G.Zoutendijk (1960), **Methods of Feasible Directions**, Elsevier, Amsterdam.

APPENDIX A GEOMETRIC PROGRAMMING

Geometric programming is generally used to optimize functions which are in the form of posynomials (positive polynomials) subject to constraints of the same type, and consists of a primal-dual pair of problems. In its original development, the generalized arithmetic-geometric mean inequality

$$\sum_{i=1}^n u_i \geq \prod_{i=1}^n (u_i/\lambda_i)^{\lambda_i} \quad (\text{A.1})$$

is employed to convert a GP primal problem to its dual where all u_i are nonnegative, and λ_i are positive weights satisfying the following normality condition

$$\sum_{i=1}^n \lambda_i = 1 \quad (\text{A.2})$$

The inequality (A.1) becomes an equality if and only if all u_i are equal.

In the present study, only the simplest form of GP problems is used which is concerned with the unconstrained minimization of a posynomial function consisting of n terms and having m variables denoted by y_j , $j=1, \dots, m$. This minimization problem constitutes the GP primal problem:

(G)

$$\text{Min } f(y) = \sum_{i=1}^n c_i \prod_{j=1}^m (y_j)^{a_{ji}} \quad (\text{A.3})$$

where $c_i > 0$ ($i=1, \dots, n$), $y_j > 0$ ($j=1, \dots, m$) and a_{ji} ($i=1, \dots, n$, $j=1, \dots, m$) are unrestricted in sign.

Corresponding to problem (G) there is a GP dual problem (DG) which consists of maximizing a dual objective function over dual variables p_i ($i=1, \dots, n$) subject to linear equality constraints.

(DG)

$$\text{Max } d(p) = \prod_{i=1}^n (c_i/p_i)^{P_i} \quad (\text{A.4})$$

$$\sum_{i=1}^n p_i = 1 \quad (\text{A.5})$$

$$\sum_{i=1}^n a_{ji} p_i = 0 \quad j=1, \dots, m \quad (\text{A.6})$$

where $p_i \geq 0$.

The relationships between the primal and dual objective functions are that

$$f(y) \geq f(y^*) = d(p^*) \geq d(p) \quad (\text{A.7})$$

in which the superscript asterisk denotes optimal values. The relationships between the primal and dual variables at the optimum are:

$$p_i^* = (c_i/d(p^*)) \prod_{j=1}^m (y_j^*)^{a_{ji}} \quad i=1, \dots, n \quad (\text{A.8})$$

Problems (G) and (DG) together with relationships (A.7) and (A.8) define the primal-dual pair of GP problems and can be rigorously proved[26].

APPENDIX B

PUBLISHED PAPERS

Appendix B1

ENTROPY DÚALS

ENTROPY DUALS

ANDREW B. TEMPLEMAN

*Department of Civil Engineering,
University of Liverpool, UK*

and

LI XINGSI

*Research Institute of Engineering Mechanics,
Dalian Institute of Technology, Dalian, China*

(Received February 1, 1985)

The convex nature of Jaynes's maximum entropy formalism implies the existence of mathematical duals. The paper derives two dual forms of entropy using classical Lagrange saddle-point conditions and posynomial geometric programming. The dual forms are shown to be equivalent under transformation. Insights into the nature of entropic processes are explored and possible interpretations of the dual entropies are given. Finally it is shown that the calculation of least biased probabilities underlying random statistical data can be done very easily using standard library subroutines for unconstrained nonlinear programming to solve the dual problem.

KEYWORDS: Entropy, Jaynes's formalism, convex duals, geometric programming, random variables, computation

1 INTRODUCTION

The maximum entropy formalism, first published in 1957 by Jaynes,¹ is recognized as a fundamental concept in information theory. It addresses the problem of determining least biased probabilities for the states of a random process and has found applications in many disciplines throughout science and technology. Much of the knowledge and research generated by the maximum entropy formalism has centred around its use as an interpretative tool for studying the nature of the information available in particular random processes. It has opened up new avenues of research, new ways of interpreting random process data.

In this paper it is shown that the convex nature of the maximum entropy formalism implies the existence of a mathematical dual form of the problem. Two ways of deriving the dual of the maximum entropy problem are presented, the first using classical Lagrangian saddle-point criteria, the second using the well-established duality theory of posynomial geometric programming. Two dual forms are derived and are shown to be equivalent under variable transformation. Consideration is then given to possible interpretations of the entropy duals in relation to the primal maximum entropy formalism. The dual forms yield additional insights into the nature of entropic processes in an information theory context and clarify some of the ambiguities in the literature.

The calculation of least biased probabilities from random data has always been hampered by the fact that the maximum entropy formalism is a rather awkward numerical tool to use in all but the simplest of examples. An additional aspect of the dual forms is that they afford a simple means of calculating such probabilities for complex problems using only standard unconstrained minimization algorithms. Examples are presented.

2 THE MAXIMUM ENTROPY FORMALISM

Consider a random process which can be described by a discrete random variable x which may take several discrete values x_r , $r = 1, \dots, R$. Define p_r , $r = 1, \dots, R$ to be the probability that x has the value x_r , $r = 1, \dots, R$, i.e. $p_r \equiv p(x = x_r)$. The maximum entropy formalism is concerned with assigning least biased values to the probabilities p_r using only information which can be inferred from the random process. According to Jaynes:¹

"The least prejudiced or biased assignment of probability is that which maximizes the (Shannon) entropy subject to the given information."

The informational entropy function defined by Shannon² is:

$$S = -K \sum_{r=1}^R p_r \ln(p_r) \quad (1)$$

in which S is the (Shannon) entropy and K is a positive constant.

Jaynes refers to the above statement as the maximum entropy *principle*, the use of which word has led to heated debate. In this paper it is referred to as the maximum entropy *formalism*, in line with much recent literature.

The maximum entropy formalism casts the problem of determining the discrete probabilities p_r into the form of an optimization problem. Equation (1) must be maximized over variables p_r , $r = 1, \dots, R$, subject to any available information on the random process. Precisely what constitutes this information, what can logically be inferred and how it can be encoded in constraint form clearly depends upon the particular random process under examination. However, one constraint is axiomatic; the normality condition,

$$\sum_{r=1}^R p_r = 1 \quad (2)$$

Maximization of Eq. (1) subject only to Eq. (2) is a simple calculation which yields uniform probabilities, $p_r = 1/R$, $r = 1, \dots, R$, and an entropy ratio $S_{\max}/K = \ln(R)$. This result is derived using no information whatsoever from the random process.

Suppose now that some information is available about the random process in the form of M expectation functions of the form,

$$\sum_{r=1}^R p_r f_{jr}(x) = E[f_j] \quad j = 1, \dots, M \quad (3)$$

for which values of $f_{jr}(x)$ and $E[f_j]$, $j = 1, \dots, M$; $r = 1, \dots, R$ are known (it is assumed that $M < R - 1$). Then the required probabilities must maximize Eq. (1) and also satisfy Eq. (2) and the M Eq. (3). This problem, named problem E1 is stated below for future reference.

Problem E1

$$\begin{aligned} \text{Maximize: } (S:K) &= - \sum_{r=1}^R p_r \ln(p_r) \\ p_r, r &= 1, \dots, R \\ \text{Subject to: } \sum_{r=1}^R p_r &= 1 \\ \sum_{r=1}^R p_r f_{jr}(x) &= E[f_j] \quad j = 1, \dots, M \end{aligned}$$

In problem E1 it is axiomatic that $p_r \geq 0, r = 1, \dots, R$.

3 DUAL FORMS OF ENTROPY MAXIMIZATION

If Lagrange multipliers $(1 + \lambda)$ for the normality constraint and $\mu_j, j = 1, \dots, M$ for the M expectation constraints are defined for problem E1 the Lagrangian function $L(\mathbf{p}, \lambda, \boldsymbol{\mu})$ has the form:

$$L(\mathbf{p}, \lambda, \boldsymbol{\mu}) = - \sum_{r=1}^R p_r \ln(p_r) + (1 + \lambda) \left(\sum_{r=1}^R p_r - 1 \right) + \sum_{j=1}^M \mu_j \left(\sum_{r=1}^R p_r f_{jr} - E[f_j] \right) \quad (4)$$

Necessary conditions for a stationary point of $L(\mathbf{p}, \lambda, \boldsymbol{\mu})$ are that

$$\frac{\partial L}{\partial p_r} = 0, \quad r = 1, \dots, R; \quad \frac{\partial L}{\partial \lambda} = 0;$$

and

$$\frac{\partial L}{\partial \mu_j} = 0, \quad j = 1, \dots, M$$

The first of these yields

$$p_r = \exp\left(\lambda + \sum_{j=1}^M \mu_j f_{jr}\right), \quad r = 1, \dots, R \quad (5)$$

The second yields the normality condition, Eq. (2), which, on substituting result (5), gives

$$\lambda = -\ln\left\{ \sum_{r=1}^R \exp\left(\sum_{j=1}^M \mu_j f_{jr}\right) \right\} \quad (6)$$

and

$$p_r = \frac{\exp(\sum_{j=1}^M \mu_j f_{jr})}{\sum_{r=1}^R \exp(\sum_{j=1}^M \mu_j f_{jr})}, \quad r = 1, \dots, R \quad (7)$$

The third necessary stationarity condition yields the M expectation constraints, Eqs. (3), which on substituting (7) have the form

$$\frac{\sum_{r=1}^R f_{jr} \exp(\sum_{j=1}^M \mu_j f_{jr})}{\sum_{r=1}^R \exp(\sum_{j=1}^M \mu_j f_{jr})} = E[f_j], \quad j = 1, \dots, M \quad (8)$$

Sufficiency conditions for the stationary point to correspond to a solution of problem E1 can be seen to be satisfied if the second derivatives of the Lagrangian function (4) are examined. Thus,

$$\frac{\partial^2 L}{\partial p_r^2} = \frac{-1}{p_r} \quad r = 1, \dots, R \quad (9)$$

$$\frac{\partial^2 L}{\partial p_r \partial p_k} = 0 \quad r = 1, \dots, R; k (\neq r) = 1, \dots, R \quad (10)$$

Relationships (9) and (10) show that the Hessian matrix of the Lagrangian function is diagonal with negative elements (since $p_r \geq 0$, $r = 1, \dots, R$). This is characteristic of a strictly concave function. Consequently it is deduced that the stationary point must correspond to a global maximizing point of problem E1.

The above analysis is well-known and has appeared frequently in the literature (see, for example, Tribus³). However, previous work has not explicitly pursued the analysis beyond this stage to construct a formal mathematical dual problem to problem E1. This will now be done. The construction of a dual problem hinges upon the Lagrangian saddle-point condition. The following theorem encapsulates the essential material.

Theorem

If problem A consisting of a maximization over variables \mathbf{x} has a Lagrangian function $L(\mathbf{x}, \boldsymbol{\alpha})$ which is strictly concave such that $L(\mathbf{x}^*, \boldsymbol{\alpha}^*) \geq L(\mathbf{x}, \boldsymbol{\alpha}^*)$, where superscript asterisk denotes an optimal value, then its dual problem, problem B, is strictly convex such that $L(\mathbf{x}^*, \boldsymbol{\alpha}^*) \leq L(\mathbf{x}^*, \boldsymbol{\alpha})$ and consists of a minimization over dual variables $\boldsymbol{\alpha}$.

The above theorem is central to the theoretical conditions for optimality and is proved in many texts (see Lasdon,⁴ McCormick⁵). The proof does not require $L(\mathbf{x}, \boldsymbol{\alpha})$ to be either continuous or differentiable.

The previous analysis has shown that problem E1 satisfies all of the conditions of problem A of the theorem and therefore that they are in correspondence. The dual problem corresponding to problem B of the theorem must therefore be such that

$$L(\mathbf{p}^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*) \leq L(\mathbf{p}^*, \boldsymbol{\lambda}, \boldsymbol{\mu})$$

The dual problem of problem E1 is then:

$$\begin{aligned} &\text{Minimize: } D(\boldsymbol{\lambda}, \boldsymbol{\mu}) = L(\mathbf{p}^*, \boldsymbol{\lambda}, \boldsymbol{\mu}) \\ &\boldsymbol{\lambda}, \boldsymbol{\mu}_j, j = 1, \dots, M \end{aligned}$$

Substituting p_r^* , $r = 1, \dots, R$ given by Eq. (5), λ given by Eq. (6) into the Lagrangian function (4) and using relationships (8) yields, after algebraic manipulation, the dual problem as:

Problem E2

$$\text{Minimize: } D(\boldsymbol{\mu}) = - \sum_{j=1}^M \mu_j E[f_j] + \ln \left\{ \sum_{r=1}^R \exp \left(\sum_{j=1}^M \mu_j f_{jr} \right) \right\}$$

$\mu_j, j = 1, \dots, M$

Before examining problem E2 and its relationship to problem E1 in detail, another way of deriving a dual problem for problem E1 using geometric programming is presented. Geometric programming was developed in the late 1960s and has become an established technique of mathematical programming. The Appendix to this paper states the mathematical forms of geometric programming as used in this paper. Duffin, Peterson and Zener⁶ give a comprehensive background to the topic and full proofs of all the duality relationships which are here used without proof.

First, a modified form of problem E1 is introduced which has a different form of the entropy ratio objective function. From problem E1

$$\begin{aligned} S \cdot K &= - \sum_{r=1}^R p_r \ln(p_r) \\ &= \sum_{r=1}^R \ln \left(\frac{1}{p_r} \right)^{p_r} \end{aligned}$$

i.e.

$$S \cdot K = \ln \left\{ \prod_{r=1}^R \left(\frac{1}{p_r} \right)^{p_r} \right\} \quad (11)$$

Taking exponentials of both sides of Eq. (11) gives

$$\exp(S \cdot K) = \prod_{r=1}^R \left(\frac{1}{p_r} \right)^{p_r} \quad (12)$$

Equation (12) replaces the objective function ($S \cdot K$) in problem E1 to give a modified maximum entropy problem (ME1).

Problem ME1

$$\text{Maximize: } G(\mathbf{p}) \equiv \exp(S \cdot K) = \prod_{r=1}^R \left(\frac{1}{p_r} \right)^{p_r}$$

$p_r, r = 1, \dots, R$

$$\text{Subject to: } \sum_{r=1}^R p_r = 1$$

$$\sum_{r=1}^R p_r \{f_{jr} - E[f_j]\} = 0 \quad j = 1, \dots, M$$

It can be noted here that taking the exponential of the entropy function in no way affects its concavity. Values of p_r^* , $r = 1, \dots, R$ which uniquely solve problem E1 must also uniquely solve problem ME1.

If problem ME1 is now compared with the pair of primal-dual geometric programming problems given in the Appendix it can immediately be seen that it corresponds very closely to the geometric programming dual problem, problem G2. Indeed the problems ME1 and G2 are identical if the following equivalences are defined. The dual function name, G, in G2 is equivalent to $\exp(S/K)$ in ME1; all the coefficients c_r in G2 must be set to unity for all r to correspond to ME1, and all the quantities a_{jr} , $j = 1, \dots, M$; $r = 1, \dots, R$ in G2 must be set to equal the quantities $\{f_{jr} - E[f_j]\}$, $j = 1, \dots, M$; $r = 1, \dots, R$, in ME1. Also, now, the dual variables p_r , $r = 1, \dots, R$ in problem G2 can be seen to be identical to the variables in problem ME1 which are the discrete probabilities.

Having established that problem ME1 corresponds to a particular case of problem G2 in which the above equivalences are made, it is now possible to use the duality relationships of geometric programming as given in the Appendix to construct a new problem, ME2 which bears the same relationship to ME1 as G1 does to G2. This new modified maximum entropy dual problem turns out to be:

Problem ME2

$$\text{Minimize: } F(\mathbf{y}) = \sum_{r=1}^R \prod_{j=1}^M y_j^{f_{jr} - E[f_j]}$$

$$y_j, j = 1, \dots, M$$

Appendix Eqs (A1) and (A2) then give the formal relationships which exist between the objective functions and variable values in problems ME1 and ME2 at the optimizing points.

This section has now established two pairs of primal-dual problems:

$$E1 \Leftrightarrow E2$$

and

$$ME1 \Leftrightarrow ME2$$

(13)

Furthermore, ME1 is obtained from E1 simply by taking the exponential of the entropy ratio objective function. Consequently it is to be expected that the two dual problems E2 and ME2 are also closely related. In fact problem ME2 can be obtained from problem E2 by putting

$$y_j \equiv \exp(\mu_j) \quad j = 1, \dots, M$$

$$F(\mathbf{y}) \equiv \exp[D(\boldsymbol{\mu})]$$

(14)

It is therefore possible to obtain the probabilities p_r , $r = 1, \dots, R$ which solve the maximum entropy problem, E1, by solving either of the dual problems, E2 or ME2,

and using the transformations and duality relationships shown below in relationships (15)

PRIMAL PROBLEMS (MAXIMIZE)		DUAL PROBLEMS (MINIMIZE)
<div style="border: 1px solid black; padding: 2px; display: inline-block; margin-bottom: 10px;">E1</div> $G(\mathbf{p}) = \exp(S/K)$ <div style="display: inline-block; vertical-align: middle; text-align: center;"> \updownarrow </div> <div style="border: 1px solid black; padding: 2px; display: inline-block; margin-top: 10px;">ME1</div>	$(S/K)_{\max}^* = D(\boldsymbol{\mu}^*)_{\min}$ \longleftrightarrow $p_r^* = \frac{\exp(\sum_{j=1}^M \mu_j^* f_{jr})}{\sum_{r=1}^R \exp(\sum_{j=1}^M \mu_j^* f_{jr})}$ $G(\mathbf{p}^*)_{\max} = F(\mathbf{y}^*)_{\min}$ \longleftrightarrow $p_r^* = \frac{\prod_{j=1}^M y_j^{*f_{jr}}}{\sum_{r=1}^R \prod_{j=1}^M y_j^{*f_{jr}}}$	<div style="border: 1px solid black; padding: 2px; display: inline-block; margin-bottom: 10px;">E2</div> $y_j = \exp(\mu_j)$ <div style="display: inline-block; vertical-align: middle; text-align: center;"> \updownarrow </div> $F(\mathbf{y}) = \exp(D(\boldsymbol{\mu}))$ <div style="border: 1px solid black; padding: 2px; display: inline-block; margin-top: 10px;">ME2</div>

(15)

4 INTERPRETATION OF THE ENTROPY DUALS

In interpreting the results expressed by the relationships (15), an important feature is that the primal and dual problems exist in different spatial domains. The primal problems E1 and ME1 are the well-known problems of entropy *maximization* in the R -space of the discrete probabilities p_r , but the dual problems E2 and ME2 are problems of function *minimization* in the M -space of the dual variables μ_j or y_j . There is no transformation in a horizontal sense in relationships (15) between a general, non-optimal point in primal space and another point in dual space. The only point at which primal and dual spatial domains meet is at the unique optimizing point where entropy reaches its constrained maximum value and the dual function reaches its minimum value, at which point corresponding primal and dual objective functions have the same numerical value. The spatial dimensionality of the primal and dual problems is interesting: the R -space of the primal problems is that of the discrete probabilities, the M -space of the dual problems is that of the available information (M is the number of statistical averages or moments which are known). It is therefore clear that whatever the dual function $D(\boldsymbol{\mu})$ (or its alternative form $F(\mathbf{y})$) represents, its is definitely not entropy.

The function $D(\boldsymbol{\mu})$ in problem E2 appears frequently in the literature but has not, to the authors' knowledge, been specifically identified as the formal convex dual of problem E1. Agmon, Alhassid and Levine⁸ introduce the exact form of $D(\boldsymbol{\mu})$ which they call a "potential function." They note that minimizing $D(\boldsymbol{\mu})$ over variables $\mu_j, j = 1, \dots, M$ produces optimal values μ_j^* which are equal to the Lagrange parameters of problem E1. Neither in that paper,⁸ nor in earlier work⁹ which uses the potential function is there any reference to the fact that S/K and $D(\boldsymbol{\mu})$ are the elements of a primal-dual pair, although all the necessary ingredients are present. Agmon *et al.*⁸ suggest only that $D(\boldsymbol{\mu})$ might be useful in a numerical algorithm to calculate the Lagrange parameters in the maximum entropy formalism.

That the general function $D(\boldsymbol{\mu})$ is not a measure of entropy is clear: precisely what $D(\boldsymbol{\mu})$ measures is not so clear. A clue is found in Jaynes's statement of the

formalism given earlier: the probabilities which maximize the entropy subject to the given information are least prejudiced or least biased assignments. This suggests that the dual of entropy maximization is prejudice or bias minimization, and that $D(\mu)$ is a direct measure of the prejudice or bias contained in the information. The term "potential function" used by Agmon *et al.*⁸ implies that $D(\mu)$ represents the usefulness, value or potential of any information that a message may contain. In order to assume nothing more than is strictly justifiable, it is necessary to extract only the minimum core of intrinsic information by minimizing the potential function.

5 NUMERICAL SOLUTION METHODS

The purely numerical problem of calculating least biased discrete probabilities p_r , $r = 1, \dots, R$ from available random data is now addressed. Assuming that expected values of statistical moments can be calculated from the data, $E[f_j]$, $j = 1, \dots, M$ and f_j , $j = 1, \dots, M$, $r = 1, \dots, R$ will be known. There are then several ways of calculating the desired probabilities. Griffeath¹⁰ solved problem E1 directly using constrained non-linear programming (a feasible directions method). An alternative method is to solve the M non-linear Eqs (8) for the M unknown Lagrange multipliers μ_j , $j = 1, \dots, M$ and substitute values of these multipliers into the R Eqs (7) to give the desired probabilities. The difficulty in this method lies in solving the M non-linear Eqs (8). One way of doing this is to write the equations in residual form and use numerical optimization to minimize the sum of the squares of the residuals. This is the basis of the methods used by Siddall and Diab¹¹ and Basu and Templeman¹² for continuous random variables although they are equally applicable to discrete random variables.

The above methods are cumbersome and require considerable programming care to avoid computational inaccuracies. They are all directed towards solving the primal problem E1. A different approach, in the spirit of Agmon *et al.*,⁸ is that of solving numerically the dual problem (E2 or ME2) and using the simple primal-dual relationship (15) to calculate the desired probabilities. This approach is discussed further here.

Both dual problems, E2 and ME2, are unconstrained non-linear minimization problems over M variables. Neither has any marked advantage over the other in respect of ease of numerical solution: problem ME2 is chosen here as it has a more concise functional form. Generally the number of variables M will be small, perhaps between one and four. This gives an advantage over Griffeath's method¹⁰ in which the number of variables is R , since in most cases $R > M$ and the computational efficiency of optimization algorithms depends largely upon the number of variables; the fewer the better. Also, unconstrained minimization problems such as ME2 are in the simplest class of all optimization problems to solve numerically, especially with few variables. Problem E1 as solved by Griffeath is a constrained problem and consequently is inherently more difficult to solve.

The dual problem ME2 has other attributes which help in its rapid solution. Firstly the function $F(y)$ is differentiable and closed-form algebraic derivatives can be specified very easily. This means that the most efficient unconstrained optimization algorithms which employ first (and sometimes second) derivatives can be used. A second important feature is that it is always possible to specify a good starting point for the numerical search process. This consists of initially setting all the variables y_j , $j = 1, \dots, M$ in problem ME2 to unity. (In problem E2 the corresponding

TABLE I
Moment data used in test examples

Test no.	$E[x]$	$E[x^2]$	$E[x^3]$	$E[x^4]$
1	4.5	--	--	--
2	4.5	22.75	--	--
3	4.5	22.75	123.0	--
4	4.5	22.75	123.0	690.0

initial values are $\mu_j = 0, j = 1, \dots, M$.) It is easy to show from relationships (15) that this starting point is equivalent to making an initial guess that the probabilities $p_r, r = 1, \dots, M$ in problem E1 are uniformly distributed. If the numerical optimization process subsequently alters the values of the variables y_j so as to minimize $F(y)$, this reflects the fact that, and the extent to which, the extra information provided by the statistical data causes the uniform probabilities to change to a different distribution. It can also be noted here that if no informational constraints (3) exist in problem E1 then $M = 0$ and problems E2 and ME2 do not exist. They only exist when information is present and can therefore be thought of as modelling the effect which that information has upon uniformly distributed probabilities. The strictly convex nature of problems E2 and ME2 ensures that the globally optimal point will be reached whatever the starting point may be.

The computational solution of the dual problem E2 or ME2 is simple and rapid. A NAG library subroutine was used for the unconstrained minimization, the remainder of the short program consisting of data input and output statements and details of the function to be minimized, its gradients and starting point. The NAG library subroutine used in the examples here was EO4DFF, a modified-Newton algorithm devised by Gill and Murray.¹³

The numerical example is based upon an imaginary die-rolling experiment using an unfair die. Test number one postulates that after many rolls the mean of the face-up value is $E[x] = 4.5$ (instead of 3.5 for a fair die). In test number two it is assumed that both the mean and variance of the face-up values are known, $E[x] = 4.5$ and $E[x^2] = 22.75$ (instead of 15.666 for a fair die). In tests three and four hypothetical values for $E[x^3]$ and $E[x^4]$ are given as shown in Table I. For each test the underlying discrete probabilities which maximize entropy are found by using EO4DFF to minimize the unconstrained entropy dual problem. Results are given in Table II.

TABLE II
Resulting discrete probabilities

Test no.	1	2	3	4
p_1	0.054354	0.067414	0.015168	0.000030
p_2	0.078772	0.081046	0.135295	0.176936
p_3	0.114161	0.105414	0.182140	0.166962
p_4	0.165447	0.148336	0.115131	0.062205
p_5	0.239774	0.225829	0.106299	0.166814
p_6	0.347491	0.371961	0.445967	0.427054
$(S K)_{\max}$	1.613581	1.609575	1.491612	1.440478

Little comment on the results is necessary. The hypothetical die which led to the expected value data is clearly grossly unfair. Probabilities and entropies are stated to six decimal places only; the computational accuracy was higher. The results of test one can be verified by hand calculation. Results for the other tests were obtained rapidly and with minimal programming effort.

6 ENTROPY DUALS FOR CONTINUOUS RANDOM VARIABLES

Dual forms similar in nature to E2 and ME2 can be derived for the case in which the random variable x is continuous-valued rather than discrete-valued. If $p(x)$ is the probability distribution function of the continuous random variable x the maximum entropy problem paralleling problem E1 has the form:

Problem CE1

$$\text{Maximize: } (S/K) = - \int p(x) \ln[p(x)] dx$$

$p(x)$

$$\text{Subject to: } \int p(x) dx = 1$$

$$\int p(x) f_j(x) dx = E[f_j] \quad j = 1, \dots, M$$

in which integration is performed over the range of variable x .

Using the same mathematical approach as in Section 3 it can be demonstrated that the dual form of CE1 is similar to problem E2:

Problem CE2

$$\text{Minimize: } D(\mu) = - \sum_{j=1}^M \mu_j E[f_j] + \ln \left\{ \int \exp \left(\sum_{j=1}^M \mu_j f_j(x) \right) dx \right\}$$

$\mu_j, j = 1, \dots, M$

Solving problem CE2 over variables μ yields $D(\mu^*)$ and μ^* from which $(S/K)_{\max}^*$ and $p^*(x)$ are found using:

$$(S/K)_{\max}^* = D(\mu^*)_{\min} \quad (16)$$

and

$$p^*(x) = \frac{\exp(\sum_{j=1}^M \mu_j f_j(x))}{\int \exp(\sum_{j=1}^M \mu_j f_j(x)) dx} \quad (17)$$

By use of the transformations (14) a more concise dual formulation which parallels problem ME2 can be obtained as:

Problem CME2

$$\text{Minimize: } F(\mathbf{y}) = \int \prod_{j=1}^M y_j^{f_j(x) - E[f_j]} dx$$

$$y_j, j = 1, \dots, M$$

The maximum entropy based probability function $p^*(x)$ can be recovered from relationships (14), (16) and (17) after solving the unconstrained minimization problem CME2.

From a numerical solution viewpoint the continuous dual problems CE2 and CME2 have all the attributes of their discrete counterparts. They are slightly more complicated in that they each involve a single integration to evaluate the dual function at each trial point in the numerical minimization process. Because of the nature of the integrand, numerical integration must be used. However, the alternative solution methods for continuous entropy maximization described in References 11 and 12, which attempt to solve problem CE1 directly require M numerical integrations at each trial point. Consequently, obtaining maximum entropy probability distributions via the dual forms CE2 or CME2 should be considerably simpler and more rapid than the alternative methods.

Finally, it may be noted that objections have been raised in the literature to the use of the continuous entropy form $(S/K) = \int p(x) \ln[p(x)] dx$ on the grounds that it is not invariant under variable transformation. Nevertheless, this is the form most often used in practical applications and this Section is therefore presented on a pragmatic basis.

7 CONCLUSIONS

A dual formulation has been developed using the Lagrangian saddle-point condition for estimating discrete and continuous probability distributions which satisfy expectation-constrained maximum entropy criteria.

It has been demonstrated that the dual entropy problems have distinct advantages in respect of their ease of use in numerical algorithms for estimating least biased probabilities. Computer programs employing dual entropy formulations are small and easy to program. They can use standard library subroutines for unconstrained minimization (and numerical integration) and are very rapid in execution.

REFERENCES

1. E. T. Jaynes, "Information Theory and Statistical Mechanics," *The Physical Review*, **106**, 620-630 and **108**, 171-190 (1957).
2. C. E. Shannon, "A Mathematical Theory of Communication," *Bell System Technical Journal*, **27**, (3), 379 (1948).
3. M. Tribus, *Rational Descriptions, Decisions and Designs*, Pergamon, New York (1969).
4. L. S. Lasdon, *Optimization Theory for Large Systems*, Macmillan (1970).
5. G. P. McCormick, *Nonlinear Programming. Theory, Algorithms, and Applications*, Wiley (1983).
6. R. J. Duffin, E. L. Peterson and C. Zener, *Geometric Programming*, Wiley (1967).
7. M. Tribus, "Information Theory as the Basis for Thermostatistics and Thermodynamics," *ASME J. Applied Mechanics*, Paper No. 60-WA-23, 1-8, March (1961).
8. N. Agmon, Y. Alhassid and R. D. Levine, "An Algorithm for Determining the Lagrange Parameters in the Maximum Entropy Formalism," in *The Maximum Entropy Formalism* (R. D. Levine and M. Tribus (Eds.)), MIT Press, 207-209 (1978).

9. R. D. Levine, "Entropy and Macroscopic Disequilibrium. II. The Information Theoretic Characterization of Markovian Relaxation Processes," *AIPh. J. Chemical Physics*, **65**, (8), 3302-3315 (1976).
10. D. S. Griffeth, "Computer Solution of the Discrete Maximum Entropy Problem," *Technometrics*, **14**, (4), 891-897 (1972).
11. J. N. Siddall and Y. Diab, "The use in Probabilistic Design of Probability Curves generated by Maximizing the Shannon Entropy Function Constrained by Moments," *ASME, Paper No. 74-WA/DE-2* (1974).
12. P. C. Basu and A. B. Templeman, "An Efficient Algorithm to generate Maximum Entropy Distributions," *Int. J. Numerical Methods in Engineering*, **20**, 1039-1055 (1984).
13. P. E. Gill and W. Murray, "Newton-type methods for unconstrained and linearly constrained optimization," *Mathematical Programming*, **28**, 311-350 (1974).

Appendix—Geometric programming⁶

Geometric programming consists of a primal-dual pair of problems and is here used in its simplest form which is concerned with the unconstrained minimization of a posynomial function (a polynomial function with strictly positive coefficients).

Consider the problem of minimizing a posynomial function consisting of R terms $r = 1, \dots, R$ and having M variables denoted by y_j , $j = 1, \dots, M$. This notation is deliberately chosen to be identical to notation already defined for maximum entropy problems to facilitate subsequent demonstrations of equivalence. This constrained minimization problem constitutes the geometric programming primal problem (G1).

Problem G1

$$\text{Minimize: } F(\mathbf{y}) = \sum_{r=1}^R c_r \prod_{j=1}^M y_j^{a_{jr}}$$

$$y_j, j = 1, \dots, M$$

$$\text{in which: } \begin{array}{ll} c_r > 0 & r = 1, \dots, R \\ y_j > 0 & j = 1, \dots, M \end{array}$$

$$a_{jr} \text{ is unrestricted in sign, } r = 1, \dots, R; j = 1, \dots, M$$

Corresponding to problem G1 there is a geometric programming dual problem, G2, which consists of maximizing a dual objective function $G(\mathbf{p})$ over dual variables p_r , $r = 1, \dots, R$, subject to linear equality constraints. Here p_r , $r = 1, \dots, R$ is used simply to name or identify dual variables. Equivalence between the dual variables and discrete probabilities p_r , $r = 1, \dots, R$ are demonstrated in the main text.

Problem G2

$$\text{Maximize: } G(\mathbf{p}) = \prod_{r=1}^R \left(\frac{c_r}{p_r} \right)^{p_r}$$

$$p_r, r = 1, \dots, R$$

$$\text{subject to: } \sum_{r=1}^R p_r = 1$$

$$\sum_{r=1}^R a_{jr} p_r = 0 \quad j = 1, \dots, M$$

$$\text{in which: } p_r \geq 0, \quad r = 1, \dots, R$$

The relationships between the primal and dual objective functions are that:

$$F(y_1, \dots, y_M) \geq F(y_1^*, \dots, y_M^*) = G(p_1^*, \dots, p_R^*) \geq G(p_1, \dots, p_R) \quad (\text{A1})$$

in which the superscript asterisk denotes optimal values. The relationships between the primal and dual variables, \mathbf{y} and \mathbf{p} , at the optimum are:

$$p_r^* = \frac{c_r}{G^*} \prod_{j=1}^M y_j^{a_{jr}} \quad r = 1, \dots, R \quad (\text{A2})$$

Problems G1 and G2 together with relationships (A1) and (A2) define geometric programming and can be rigorously proved.⁶

Appendix B2

A MAXIMUM ENTROPY APPROACH TO
CONSTRAINED NON-LINEAR PROGRAMMING

A MAXIMUM ENTROPY APPROACH TO CONSTRAINED NON-LINEAR PROGRAMMING

ANDREW B. TEMPLEMAN

*Department of Civil Engineering, University of Liverpool,
P.O. Box 147, Liverpool, UK*

and

LI XINGSI

*Research Institute of Engineering Mechanics,
Dalian Institute of Technology, Dalian, China*

(Received 12 January 1987)

The paper explores the use of the Shannon (informational) entropy measure and Jaynes's maximum entropy formalism in the solution of constrained non-linear programming problems. Through a surrogate constraint approach an entropy based update formula for the surrogate multipliers is derived. A numerical example of the method is presented. Some information-theoretic interpretations of mathematical programming are explored. Finally, through the use of surrogate duals the method is extended into an entropy augmented Lagrangean formulation.

KEYWORDS: Entropy, information theory, mathematical programming, surrogate constraints, duality.

1 INTRODUCTION

This paper presents some new insights on methods for solving the general constrained non-linear programming problem:

Problem A

$$\begin{array}{ll} \text{Minimize} & f(\mathbf{x}) \\ & x_i, i = 1, \dots, N \end{array} \quad (1)$$

$$\text{Subject to} \quad g_j(\mathbf{x}) \leq 0 \quad j = 1, \dots, M \quad (2)$$

Problem A has many and widespread applications, and literature on solution methods for its many forms and variants is well known. Engineering design applications usually involve non-linear functions f and/or g and almost invariably have active constraints at the optimum. This is the case to be studied here.

The present work is part of a research program exploring a possible information-theoretic basis for a wide variety of engineering synthesis and analysis problems. The research is motivated by subjective appreciations on the part of the authors from two directions. These are, firstly, that the methods of engineering optimization currently available all originated in the 1960's and early 1970's and seem to have entered a

diminishing returns phase in respect of further research potential. Some radically different directions and new approaches are needed for the further development of engineering optimization techniques. Secondly, the authors believe that information theory concepts such as the Shannon measure of uncertainty and the Jaynes maximum entropy formalism are immensely powerful and fundamental, and may have far wider significant applications than research to date has shown.

The paper is essentially exploratory. At present the methods of mathematical programming contain no information-theoretic basis. They all view optimization problems in terms of a topological domain defined deterministically by function hypersurfaces. Information theory appears to be incompatible with this as it is essentially concerned with probabilities. The current topological, deterministic view is, however, only an interpretation of mathematical programming. This paper is a first step towards providing a possible alternative interpretation couched in probabilistic, information-theoretic terms. Problem A has been chosen for study because the authors have found that it is possible to develop a solution method for it which uses the Shannon/Jaynes entropy formalism. Consequently it is possible to construct an information-theoretic, probabilistic interpretation of problem A. The paper describes the new solution method for problem A and its interpretation and explores some new insights which the information theory approach yields on some other aspects of mathematical programming.

2 SURROGATE NON-LINEAR PROGRAMMING

In problem A, \mathbf{x} denotes the vector of real, continuous-valued variables $x_i, i = 1, \dots, N$, and the constraint functions $g_j, j = 1, \dots, M$ comprise constraint vector \mathbf{g} . Problem A has an equivalent *surrogate* form:

Problem B

$$\begin{aligned} & \text{Minimize}_{\mathbf{x}, i=1, \dots, N} && f(\mathbf{x}) \\ & \text{Subject to} && \sum_{j=1}^M \lambda_j g_j(\mathbf{x}) = 0 \end{aligned} \quad (3)$$

In problem B, the $\lambda_j, j = 1, \dots, M$ are non-negative multipliers, termed surrogate multipliers, forming a vector λ . We are here concerned with solving problem A for instances in which at least one of the constraints is active at the optimum. Consequently the surrogate constraint (3) in problem B, which in its most general form should be written as ≤ 0 , must be satisfied as an equality at the optimum and is stated here as a strict equality. Also, since its right-hand side is zero, the surrogate multipliers λ may be assumed to be normalized without loss of generality, i.e.

$$\sum_{j=1}^M \lambda_j = 1 \quad (4)$$

The desired solution \mathbf{x}^* of problem A will here be sought indirectly through a sequence of solutions of problem B. This approach assumes, therefore, that problems A and B are equivalent at the solution point; specifically that a set of multipliers λ^* exists and can be found such that \mathbf{x}^* which solves problem B with λ^* also solves

problem A. Previous work by Everett,¹ Brooks and Geoffrion,² Gould,³ and Greenberg and Pierskalla⁴ have studied this assumption, establishing conditions on its validity. The proof is not repeated here but is similar to existence and optimality proofs for Lagrange multipliers in non-linear programming. An essential condition to be satisfied is that $(\mathbf{x}^*, \lambda^*, \alpha^*)$ should be a saddle point of the Lagrangean, L_B of problem B. L_B has the form

$$L_B(\mathbf{x}, \lambda, \alpha) = f(\mathbf{x}) + \alpha \sum_{j=1}^M \lambda_j g_j(\mathbf{x}) \tag{5}$$

where α is the Lagrange multiplier associated with constraint (3) in problem B. The saddle point $(\mathbf{x}^*, \lambda^*, \alpha^*)$ must be such that

$$L_B(\mathbf{x}, \lambda^*, \alpha^*) \geq L_B(\mathbf{x}^*, \lambda^*, \alpha^*) \geq L_B(\mathbf{x}^*, \lambda, \alpha) \tag{6}$$

In (6) the left-hand inequality implies minimization over variables \mathbf{x} for specified λ and α ; the right-hand inequality implies maximization over λ and α for specified \mathbf{x} . In fact \mathbf{x} , λ and α are functionally related, for example through the stationarity of L_B with respect to \mathbf{x} :

$$\frac{\partial L_B}{\partial x_i}(\mathbf{x}, \lambda, \alpha) = \frac{\partial}{\partial x_i} f(\mathbf{x}) + \alpha \sum_{j=1}^M \lambda_j \frac{\partial}{\partial x_i} g_j(\mathbf{x}) = 0 \quad i = 1, \dots, N \tag{7}$$

Thus the saddle point (6) represents a duality relationship which can be generally expressed as

$$\text{Min}_{\mathbf{x}} \phi(\mathbf{x}) \geq \text{Max}_{\lambda, \alpha} \psi(\lambda, \alpha) \tag{8}$$

In a few special cases, for example when $f(\mathbf{x})$ and $g(\mathbf{x})$ are convex separable functions, it is possible to construct functions ϕ and/or ψ in (8) and the primal-dual nature of the saddle point then leads to practical solution algorithms. In most cases, however, (8) is not of direct value: ϕ or ψ in (8) cannot be established explicitly since relationships such as (7) relate \mathbf{x} , λ and α implicitly. This more general case is addressed here.

The saddle point condition (6) is written in terms of the Lagrangean, L_B , of problem B. This is notationally convenient. It may be satisfied, however, by iterative means using problem B itself, with alternative iterations in \mathbf{x} -space and λ -space. A typical scheme is as follows. An initial set of multipliers λ^0 is chosen and problem B is solved to yield corresponding values of \mathbf{x}^0 (by minimization, corresponding to the left-hand inequality in (6)). The multipliers are then updated to λ^1 (by a maximization process for fixed \mathbf{x}^0 , corresponding to the right-hand inequality in (6)) and problem B is solved again to give \mathbf{x}^1 . The process is repeated until the sequence $(\lambda^0, \mathbf{x}^0)$, $(\lambda^1, \mathbf{x}^1)$, ..., $(\lambda^k, \mathbf{x}^k)$, ... converges upon a solution of problem B, and hence also of problem A, at $(\lambda^*, \mathbf{x}^*)$.

The major snag in this iterative scheme lies in updating the multipliers λ . It is clear from (6) that the updates should satisfy some maximization criterion, but there is no explicit criterion to govern this process. Surrogate algorithms have hitherto based multiplier updates upon gradient information only. Thus, if at iteration k of the algorithm \mathbf{x}^k has been found by solving problem B for multipliers λ^k , it is evident that updated multipliers λ^{k+1} for the next iteration should be such that

$$\lambda_j^{k+1} \begin{cases} > \lambda_j^k & \text{if } g_j(\mathbf{x}^k) > 0 \\ = \lambda_j^k & \text{if } g_j(\mathbf{x}^k) = 0 \\ < \lambda_j^k & \text{if } g_j(\mathbf{x}^k) = 0 \end{cases} \tag{9}$$

with

$$\sum_{j=1}^M \lambda_j^{k+1} = 1 \quad (10)$$

It is easy to see whether each multiplier should increase or decrease in value but no information is available to determine the magnitudes of the changes. This paper presents a method for determining an efficient update formula for the multipliers λ based upon the maximum entropy formalism.

3 THE MAXIMUM ENTROPY FORMALISM

The maximum entropy formalism published by Jaynes⁵ in 1957 is a fundamental concept in information theory. In general terms it is concerned with establishing what logical, unbiased inferences can be drawn from available information. Its use, however, has been limited to only a few areas of research specifically concerned with statistical inference, typically in establishing prior probabilities in a range of practical applications in science and engineering. In such applications the maximum entropy formalism is used to determine the probabilities underlying a random process from any available statistical data about the process. The prior probabilities thus established may then be incorporated into decision-making or control strategies for the process. References [6] to [8] typify this use of the maximum entropy formalism.

Another related area of research has been the use of the formalism to justify and validate some of the implicit assumptions and axioms in the mathematical modelling of technological processes.⁹ The use of the formalism in this way can sometimes reveal additional insights and assumptions which may initially be made, thus sharpening the results of subsequent analysis.

Exploration of the axioms of science and engineering leads into what is perhaps the most potentially important, and certainly is the most controversial, area of application of the maximum entropy formalism. This concerns the question of whether informational entropy is axiomatic to the accepted principles and laws of mechanics and physics. Jaynes consistently refers to informational entropy maximization as a *principle*, implying that all else is consequent upon its fundamental nature. This view has many adherents including Tribus¹⁰ who has attempted to show that the principles of thermostatics and thermodynamics may be derived from the informational maximum entropy principle. The converse view, also widely held, is that informational maximum entropy is merely a very useful *formalism*—a convenient means of understanding and representing a physical principle in mathematical terms. Whatever judgement the future might make, it is true to say that informational entropy maximization is a very powerful concept which is as yet relatively unexplored. The following exposition is couched in probabilistic terms for conciseness and convenience although a slightly different interpretation of it will later be made.

Consider a random process which can be described by a discrete random variable x which can have several discrete values x_r , $r = 1, \dots, R$. Define p_r , $r = 1, \dots, R$ to be the probability that x has the value x_r , $r = 1, \dots, R$, i.e. $p_r \equiv p(x = x_r)$. The maximum entropy formalism is concerned with assigning least biased values to the probabilities p_r using only information which can be inferred from the random process itself. According to Jaynes:⁵

“The least prejudiced or biased assignment of probability is that which maximizes the (Shannon) entropy subject to the given information.”

The informational entropy function defined by Shannon¹¹ is

$$S = -K \sum_{r=1}^R p_r \ln(p_r) \tag{11}$$

in which S is the (Shannon) entropy and K is a positive constant.

The maximum entropy formalism casts the problem of determining the discrete probabilities p_r into the form of an optimization problem. Equation (11) must be maximized over variables $p_r, r = 1, \dots, R$, subject to any available information on the random process. Precisely what constitutes this information, what can logically be inferred and how it can be encoded in constraint form depends upon the particular random process under examination. However, one constraint is axiomatic; the normality condition,

$$\sum_{r=1}^R p_r = 1 \tag{12}$$

Maximization of Eq. (11) subject only to (12) is simple and yields uniform probabilities, $p_r = (1/R), r = 1, \dots, R$ and an entropy $S = K \ln(R)$. Thus far no information from the random process has been used and the formalism has shown that in the absence of information equal probability must be assigned to all possible outcomes.

Suppose, now, that some information is available about the random process in the form of T expectation functions of the form

$$\sum_{r=1}^R p_r h_t(x_r) = E[h_t(x)] \quad t = 1, \dots, T \tag{13}$$

for which values of $h_t(x_r)$ and $E[h_t(x)]$ are known for $r = 1, \dots, R$ and $t = 1, \dots, T$. $E[\cdot]$ denotes the expected value of the function \cdot . It is assumed that $T < R - 1$. Then the desired least-biased probabilities must maximize Eq. (11) and also satisfy Eqs (12) and (13), i.e.

Problem C

$$\text{Maximize}_{p_r, r=1, \dots, R} S = -K \sum_{r=1}^R p_r \ln(p_r)$$

$$\text{Subject to } \sum_{r=1}^R p_r = 1$$

$$\sum_{r=1}^R p_r h_t(x_r) = E[h_t(x)] \quad t = 1, \dots, T$$

In problem C it is axiomatic that $p_r \geq 0, r = 1, \dots, R$.

The solution of problem C can be shown to be (see Tribus¹⁰)

$$p_r = \exp \left[\sum_{t=1}^T v_t h_t(x_r) \right] / \left\{ \sum_{r=1}^R \exp \left[\sum_{t=1}^T v_t h_t(x_r) \right] \right\} \quad r = 1, \dots, R \tag{14}$$

in which $v_t, t = 1, \dots, T$ are the Lagrange multipliers associated with the expectation constraints (13) in problem C.

The T Lagrange multipliers \mathbf{v} can, in theory, be found by substituting result (14) into the T expectation constraints (13) and solving uniquely for \mathbf{v} . Solution is not easy as these equations are non-linear and can usually only be solved numerically. Templeman and Li^{1,2} have shown that the Lagrange multipliers \mathbf{v} may be found more easily by solving problem D which is the convex dual of problem C.

Problem D

$$\text{Minimize } D(\mathbf{v}) = - \sum_{t=1}^T v_t E[h_t(\mathbf{x})] + \ln \left\{ \sum_{r=1}^R \exp \left[\sum_{t=1}^T v_t h_t(x_r) \right] \right\}$$

Problem D is an unconstrained minimization over the variables \mathbf{v} . At the solution point, $D(\mathbf{v}^*) = S(\mathbf{p}^*)/K$.

This concludes the brief survey of the maximum entropy formalism. We are now in a position to re-examine the problem of how best to update the surrogate multipliers in an iterative solution scheme for problem B.

4 ENTROPY-BASED SURROGATE MULTIPLIER UPDATES

Suppose that we have somehow assigned values to the surrogate multiplier set λ^k and have solved problem B with λ^k over variables \mathbf{x} to yield \mathbf{x}^k . Suppose that the constraint function values $g_j(\mathbf{x}^k)$, $j = 1, \dots, M$ have been calculated and that at least one of the set \mathbf{g}^k is positive (indicating a constraint violation in problem A, and consequently that $\mathbf{x}^k \neq \mathbf{x}^*$). We now wish to continue iterations by updating λ^k to λ^{k+1} and solving problem B again with λ^{k+1} over variables \mathbf{x} . How can we update λ^k to λ^{k+1} ?

The new multipliers λ^{k+1} must clearly satisfy the normality condition (10) and the surrogate constraint in problem B. Thus,

$$\sum_{j=1}^M \lambda_j^{k+1} = 1 \quad (10)$$

$$\sum_{j=1}^M \lambda_j^{k+1} g_j(\mathbf{x}^{k+1}) = 0 \quad (15)$$

These two linear equations are generally insufficient to yield a unique assignment of the M multipliers λ^{k+1} and no explicit objective function in λ^{k+1} is available to provide such an assignment. Also, Eq. (15) requires values for $g_j(\mathbf{x}^{k+1})$, $j = 1, \dots, M$ and these are not yet known. The best current estimates for $\mathbf{g}(\mathbf{x}^{k+1})$ are the values of $\mathbf{g}(\mathbf{x}^k)$ which are available and must be used in their place. Consequently, Eq. (15) is modified to

$$\sum_{j=1}^M \lambda_j^{k+1} g_j(\mathbf{x}^k) = \epsilon \quad (16)$$

where ϵ represents the unknown error introduced by using $\mathbf{g}(\mathbf{x}^k)$ in place of $\mathbf{g}(\mathbf{x}^{k+1})$. Bearing in mind that we wish λ_j^{k+1} and λ_j^k to satisfy the relationship (9) it may be deduced that the error term ϵ in (16) must be positive. Furthermore we would expect ϵ

to approach zero as the sequence of iterative results $\mathbf{x}^k, \mathbf{x}^{k+1}, \dots$ approaches \mathbf{x}^* . We are therefore faced with the following assignment problem:

$$\begin{aligned} \text{Assign} \quad & \lambda_j^{k+1} \quad j = 1, \dots, M \\ \text{Subject to} \quad & \sum_{j=1}^M \lambda_j^{k+1} = 1 \end{aligned} \tag{10}$$

$$\sum_{j=1}^M \lambda_j^{k+1} g_j(\mathbf{x}^k) = \epsilon \tag{16}$$

with $\epsilon \rightarrow 0^+$.

Only the cases for $M \geq 2$ are of interest (surrogation has no meaning for a single constraint, $M = 1$). For the special case $M = 2$ equations (10) and (16) yield a unique assignment:

$$\lambda_1^{k+1} = [\epsilon - g_2(\mathbf{x}^k)]/[g_1(\mathbf{x}^k) - g_2(\mathbf{x}^k)] \tag{17a}$$

$$\lambda_2^{k+1} = [g_1(\mathbf{x}^k) - \epsilon]/[g_1(\mathbf{x}^k) - g_2(\mathbf{x}^k)] \tag{17b}$$

which depends upon the unknown error term ϵ . For the general case $M > 2$ (10) and (16) are insufficient to yield a unique assignment λ^{k+1} , yet one is required. In the absence of any explicit criterion of λ -maximization (necessary to satisfy the saddle point condition (6)) which would lead to a unique assignment, an artificial criterion must be imposed. However, on the basis of the information available, there is no logical justification for a criterion which unduly favours one specific assignment rather than another. The artificial criterion should correspond to a maximization process in λ -space, assignments resulting from it should satisfy (10) and (16) and should be in accordance with the directional requirements (9), but otherwise the criterion should introduce minimum bias or prejudice into the assignment λ^{k+1} . Hence, in the spirit of the maximum entropy formalism, it is entirely logical to maximize the entropy of the multipliers λ^{k+1} . The unsolvable assignment problem is therefore replaced by the solveable problem E:

Problem E

$$\text{Maximize}_{\lambda_j^{k+1}, j=1, \dots, M} \quad S(\lambda)/K = - \sum_{j=1}^M \lambda_j^{k+1} \ln(\lambda_j^{k+1}) \tag{18}$$

$$\text{Subject to} \quad \sum_{j=1}^M \lambda_j^{k+1} = 1 \tag{10}$$

$$\sum_{j=1}^M \lambda_j^{k+1} g_j(\mathbf{x}^k) = \epsilon \tag{16}$$

Problem E is similar in form to problem C, the classical form of the maximum entropy formalism. The surrogate multipliers λ^{k+1} play the role of the discrete probabilities \mathbf{p} . Equation (16) has the form of a single expectation constraint (13) with an expected value of ϵ . By virtue of (14), the solution of problem E gives the following assignment, λ^{k+1} :

$$\lambda_j^{k+1} = \exp[\beta g_j(\mathbf{x}^k)] / \sum_{j=1}^M \exp[\beta g_j(\mathbf{x}^k)] \quad j = 1, \dots, M \tag{19}$$

in which β is the Lagrange multiplier associated with constraint (16) in problem E. Equation (19) is proposed for use as the updating relationship for the multipliers in the iterative algorithm described in Section 2 of this paper.

The Lagrange multiplier β in (19) is of course unknown but is closely related to ϵ . Substituting (19) into (16) gives

$$\sum_{j=1}^M g_j(\mathbf{x}^k) \exp[\beta g_j(\mathbf{x}^k)] / \sum_{j=1}^M \exp[\beta g_j(\mathbf{x}^k)] = \epsilon \quad (20)$$

However, ϵ is also unknown but should approach zero from the positive side as iterations approach the optimum, as noted earlier. β may therefore be considered as a control parameter whose value should be chosen at each iteration such that ϵ displays the behaviour desired. It may be deduced from (20) that for ϵ to display the required behaviour β must be chosen positive and increasingly large at each iteration. The reasoning is as follows: each constraint in problem A contributes a term to the numerator summation and a term to the denominator summation in (20). Constraints in problem A which are slack as the optimum is approached may in theory be deleted from the problem so in (20) their contributed terms should approach zero as ϵ approaches zero. For such a slack constraint $g_j(\mathbf{x}^k)$ must be negative and the terms corresponding to such a constraint only approach zero if β is positive and increasingly large. If terms in (20) corresponding to active constraints are examined with this β it can be seen that numerator terms approach zero and denominator terms are positive. Hence ϵ approaches zero from above as positive β increases.

In the special case of problem A with two constraints ($M = 2$) Eqs (17) represent a unique updating formula for λ_1 and λ_2 quite independently of the maximum entropy formalism. It may be noted that for $M = 2$ the entropy-based update formula (19) yields the same results (17) after solving (20) for β , substituting in (19) and simplification. This is not surprising, but it does permit the λ update formula (19) to be viewed as a least-biased extension of the unique result (17) for $M = 2$ into the more general realm of $M > 2$.

This concludes the derivation of an entropy-based update formula for the surrogate multipliers. Problem solving may use the algorithm described in Section 2 with alternative λ -space and \mathbf{x} -space iterations. Iterations in λ -space use (19) with a sequence of user-specified, increasing values of β .

5 NUMERICAL EXAMPLE

The example is taken from Beightler and Phillips¹³ and was chosen because an analytical solution for the \mathbf{x} -phase minimization is available for any set of values of λ . Consequently, the example focusses attention on the λ -phase iterations which are the subject of this paper. The example is:

$$\begin{aligned} \text{Minimize } & f(\mathbf{x}) = (x_1 x_2 x_3)^{-1} \\ \text{Subject to } & g_1 \equiv 2x_1 + x_2 + 3x_3 - 1 \leq 0 \\ & g_2 \equiv x_1 + x_2 + x_3 - 1 \leq 0 \\ & g_3 \equiv x_1 + 3x_2 + 2x_3 - 1 \leq 0 \\ & x_1, x_2, x_3 > 0 \end{aligned}$$

Table 1

<i>k</i>	β	λ_1	λ_2	λ_3	x_1	x_2	x_3	θ_1	θ_2	θ_3	<i>f</i>
0	—	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	0.250	0.200	0.167	0.200	-0.383	0.183	120.00
1	2.718	0.463	0.095	0.442	0.228	0.177	0.141	0.056	-0.454	0.041	175.74
2	7.389	0.521	0.012	0.467	0.219	0.172	0.133	0.009	-0.476	0.001	199.61
3	20.086	0.540	0.000	0.460	0.216	0.174	0.131	-0.001	-0.479	0.000	203.11

The surrogate constraint (3) becomes

$$\lambda_1(2x_1 + x_2 + 3x_3 - 1) + \lambda_2(x_1 + x_2 + x_3 - 1) + \lambda_3(x_1 + 3x_2 + 2x_3 - 1) = 0$$

which rearranges to

$$(2\lambda_1 + \lambda_2 + \lambda_3)x_1 + (\lambda_1 + \lambda_2 + 3\lambda_3)x_2 + (3\lambda_1 + \lambda_2 + 2\lambda_3)x_3 - 1 = 0$$

For given λ , minimizing $f(x)$ subject to this constraint is a simple geometric programming problem yielding the result:

$$\begin{aligned} x_1^* &= \frac{1}{3}(2\lambda_1 + \lambda_2 + \lambda_3) \\ x_2^* &= \frac{1}{3}(\lambda_1 + \lambda_2 + 3\lambda_3) \\ x_3^* &= \frac{1}{3}(3\lambda_1 + \lambda_2 + 2\lambda_3) \end{aligned}$$

A sequence of increasing values of β must be specified for λ iterations. The main thrust of this paper is not towards algorithmic implementation or fine-tuning, so the sequence chosen is not necessarily the best one but it does demonstrate how the method works. β was specified to be e^k where k is the iteration number. Table 1 gives the iterative results.

Iterations commence by setting $\lambda_1 = \lambda_2 = \lambda_3 = \frac{1}{3}$. Note that the Jaynes maximum entropy formalism generates this starting set for λ : in the absence of any extra information about the problem the least-biased assumption we can make is that all three constraints are equally weighted. The rest of Table 1 then follows iteratively. The solution in the bottom row of Table 1 is as exact as the three decimal place accuracy permits. Convergence is very fast.

6 DISCUSSION AND INTERPRETATION

The example confirms that entropy based updates for the surrogate multipliers lead to a fast and accurate solution of problem A. It is freely admitted that the example is a relatively simple and convex problem, and that the convergence rate is dependent upon the sequence of increasing values specified for β . Other sequences than that used above have been tested and all converged, with a final β of approximately 20 to 30. The entropy maximization basis of the method appears, therefore, to be quantitatively sound.

One feature of the method which may be noted here is that updates for λ do not require any gradient evaluations or first-order information. In order to predict λ^{k+1} for the $(k + 1)$ th iteration, only values of the constraint functions themselves, $g_j(x^k)$, $j = 1, \dots, M$, evaluated in the previous iteration are required. This feature leads conveniently into a physical interpretation of the surrogate multipliers as probabilities.

In Section 3, the maximum entropy formalism was developed and stated in a probabilistic form, yet the application described subsequently was to a problem which is conventionally and almost universally interpreted and solved as a deterministic mathematical programming problem. The use of a probabilistic method in a deterministic context is suspect unless it can be justified. We now address this problem and argue that the deterministic interpretation of problems A and B is not the only possible interpretation. Problems A and B can be explained in a probabilistic context in which informational entropy maximization is both natural and justified.

To do this a non-frequentist view of probability is taken. The probability p_r of some event r represents a measure of certainty or confidence in event r . High p_r in the range $0 \leq p_r \leq 1$ indicates high certainty. This removes any connotations of repeated experiments and frequency of occurrence of event r from the interpretation.

The solution process is set in a probabilistic context by considering Problem A as posed initially and estimating what level of certainty (probability) should be assigned to the event that each constraint is active at the problem solution. Denoting these probabilities by $\lambda_j, j = 1, \dots, M$, it is known (from the definition of the problem) that at least one of the constraints must be active, so (4) must hold. However, in the absence of any numerical information about problem A, there is no logical reason to assign a higher certainty to any one constraint rather than another. All the constraints are equally likely to be active at the optimum, thus the probability assignments $\lambda_j^0 = 1/M, j = 1, \dots, M$ must be made. Although this result is intuitively obvious, it is also a direct consequence of solving the maximum entropy problem

$$\begin{aligned} \text{Maximize } (S/K) &= - \sum_{j=1}^M \lambda_j \ln(\lambda_j) \\ \lambda_j, j=1, \dots, M & \\ \text{Subject to } \sum_{j=1}^M \lambda_j &= 1; \lambda_j \geq 0 \text{ for all } j \end{aligned} \quad (21)$$

as noted earlier. This λ^0 now permits the postulation of a probabilistic optimality condition for problem A: the solution will be found at a point at which all constraints are equally likely to be active. This condition may be expressed in expectation form (13): with probabilities λ^0 the expected value of all constraints $g(\mathbf{x})$ must be zero. Thus

$$\sum_{j=1}^M \lambda_j^0 g_j(\mathbf{x}) = 0 \quad (22)$$

Now a vector \mathbf{x}^0 must be sought which satisfies (22) and also minimizes $f(\mathbf{x})$. This corresponds to problem B with fixed surrogate multipliers (probabilities) λ^0 . Solving problem B with λ^0 therefore represents solving a probabilistic approximation to problem A by satisfying the most likely and least biased constraint optimality condition which could be constructed using available information.

Generally \mathbf{x}^0 will not correspond to \mathbf{x}^* , the solution of problem A, because the optimality condition (22) is not usually exact. The λ^0 contained in (22) are only best currently available estimates which are not necessarily accurate yet. \mathbf{x}^* will be located only when each of the terms $\lambda_j g_j(\mathbf{x})$ in the summation in (3) is zero. This is the complementary slackness condition which, interpreted probabilistically, states that for a strictly positive probability that a constraint is active at the solution, the constraint function must have value zero, and if the constraint function is non-zero the probability that it is optimally active must be zero.

The current solution point \mathbf{x}^0 will usually satisfy (22) with $g_j(\mathbf{x}^0) \neq 0, j = 1, \dots, M$. However, these values of $\mathbf{g}(\mathbf{x}^0)$ give new information about problem A which can be used in the maximum entropy formalism to improve the probability assignments λ^0 to λ^1 . It is clear that if $g_j(\mathbf{x}^0) > 0$ for some j then this should increase our certainty that constraint j will be active at the optimum, and vice versa, thus

$$\left. \begin{array}{l} \lambda_j^1 > \lambda_j^0 \text{ if } g_j(\mathbf{x}^0) > 0 \\ \lambda_j^1 < \lambda_j^0 \text{ if } g_j(\mathbf{x}^0) < 0 \end{array} \right\} j = 1, \dots, M \quad (23)$$

which, it can be noted, is the form of (9). This can be codified in expectation form as follows. Equation (22) must be satisfied exactly at \mathbf{x}^0 . Thus

$$\sum_{j=1}^M \lambda_j^0 g_j(\mathbf{x}^0) = 0 \quad (24)$$

In view of (23), it can be deduced that

$$\sum_{j=1}^M \lambda_j^1 g_j(\mathbf{x}^0) > \sum_{j=1}^M \lambda_j^0 g_j(\mathbf{x}^0) \quad (25)$$

Thus we can write

$$\sum_{j=1}^M \lambda_j^1 g_j(\mathbf{x}^0) = \epsilon \quad (26)$$

where ϵ is the expected value of the left-hand side of inequality (25). ϵ should logically be small, positive, and should tend towards zero as our certainty (represented by the sequence $\lambda^0, \lambda^1, \lambda^2, \dots$) increases. Equation (26) therefore encodes the extra information provided by the trial \mathbf{x}^0 . It may be appended as an extra constraint to (21) which can be solved to yield least biased estimates of constraint probabilities λ^1 .

Using these new certainty estimates λ^1 , a new probabilistic constraint optimality criterion (Eq. (22) with λ^1 replacing λ^0) can be formed. Solving problem B with λ^1 locates a new point \mathbf{x}^1 which satisfies the constraint criterion and minimizes $f(\mathbf{x})$. This yields new information on constraint function values $\mathbf{g}(\mathbf{x}^1)$ which can be encoded into a form similar to (26), and the iterations continue thus until λ^* and \mathbf{x}^* are found.

The above has shown that it is possible to interpret the iterative surrogate solution method in a probabilistic context, and to justify the use of the maximum entropy formalism. In the context of surrogate constraints several authors including Glover^{14,15} have used the concept of surrogate constraint strength. A surrogate constraint (3) with λ^m is defined to be stronger than a surrogate constraint with λ^n if $f(\mathbf{x}^m) > f(\mathbf{x}^n)$, where \mathbf{x}^m and \mathbf{x}^n both solve problem B with λ^m and λ^n respectively. The results in Table 1 therefore show iterations of increasing surrogate constraint strength, culminating in a solution to problem A when surrogate constraint strength is maximum. Since the λ updates given by Eq. (19) maximize the entropy of the available constraint information, this suggests that constraint strength and constraint entropy are synonymous. In this light, the approach used in this paper to solve problem A can be generally expressed as:

Minimize the objective function over \mathbf{x} such that the constraint entropy is maximized over λ .

Viewed in this context, the entropy based surrogate method may be extended and developed further into alternative functional representations and models. These extensions are now addressed.

7 EXTENSIONS

The entropy based surrogate method of the previous sections is useful in allowing a probabilistic, information-based interpretation to be developed for constrained non-linear optimization: each \mathbf{x} -space minimization represents an attempt to satisfy a probabilistic constraint optimality criterion (3) whilst minimizing f . The result of each attempt provides new information on constraint values \mathbf{g} which can be used to modify the constraint activity probabilities λ in a least biased way so that an improved constraint optimality criterion (3) can be set up for the next iteration.

Having justified the approach and having established the validity of using the Shannon entropy/Jaynes formalism in a mathematical programming context, we now turn to extensions and refinements of the formulations. A logical extension of the two-phase, \mathbf{x} -space/ λ -space iteration scheme is to examine whether the phases could somehow be combined together. To do this we recognize that the two-phase scheme is essentially trying to satisfy the Lagrangean saddle point condition of the surrogate problem B. This saddle point condition is a surrogate duality relationship which establishes a surrogate primal (\mathbf{x} -space) and a surrogate dual (λ -space) formulation of problem B (inequality (8)). To develop a single-phase method we must re-examine the surrogate duality relationship. We therefore return to problem B.

If the normality condition (4) and the λ non-negativity condition are formally incorporated into problem B (their earlier omission merely simplified the treatment and did not affect its generality) the full form of problem B becomes:

Problem B'

$$\begin{aligned} & \text{Minimize } f(\mathbf{x}) \\ & \mathbf{x}_i, i = 1, \dots, N \\ & \text{Subject to } \sum_{j=1}^M \lambda_j g_j(\mathbf{x}) = 0 \\ & \sum_{j=1}^M \lambda_j = 1 \\ & \lambda_j \geq 0 \quad j = 1, \dots, M \end{aligned}$$

and its Lagrangean is

$$L'_B = f(\mathbf{x}) + \alpha \sum_{j=1}^M \lambda_j g_j(\mathbf{x}) + \mu \left(\sum_{j=1}^M \lambda_j - 1 \right); \lambda_j \geq 0 \quad j = 1, \dots, M \quad (27)$$

The stationarity of L'_B with respect to $\mathbf{x}_i, i = 1, \dots, N$ leads to Eq. (7), as before. Equation (7) relates \mathbf{x} and λ implicitly and so cannot be used to express $\mathbf{x}(\lambda)$ and $\lambda(\mathbf{x})$ which are necessary to set up the surrogate primal and dual problems expressed explicitly by (8).

However, in the two-phase method described earlier we overcame the absence of an explicit surrogate dual objective function by introducing the Shannon entropy as a means of forcing iterations towards the saddle point. We can use the same approach

here if we augment the Lagrangean L'_B with an entropy term. Equation (27) then becomes

$$L''_B = f(\mathbf{x}) + \alpha \sum_{j=1}^M \lambda_j g_j(\mathbf{x}) + \mu \left(\sum_{j=1}^M \lambda_j - 1 \right) - \frac{1}{\rho} \sum_{j=1}^M \lambda_j \ln(\lambda_j) \quad (28)$$

The negative sign of the entropy term in (28) is necessary because of the negativity of the summation. ρ is a positive constant. Note that the λ non-negativity condition in (27) is no longer necessary as the entropy term ensures it.

Stationarity of L''_B with respect to $x_i, i = 1, \dots, N$ leads to (7), as before, but stationarity with respect to $\lambda_k, k = 1, \dots, M$ gives

$$\frac{\partial L''_B}{\partial \lambda_k} = \alpha g_k(\mathbf{x}) + \mu - \frac{1}{\rho} [\ln(\lambda_k) + 1] = 0 \quad k = 1, \dots, M \quad (29)$$

Equation (29) may be used to separate out the primal and dual variables explicitly. Solving (29) for $\lambda_k, k = 1, \dots, M$ gives

$$\lambda_k = \exp[\rho(\alpha g_k(\mathbf{x}) + \mu) - 1] \quad k = 1, \dots, M \quad (30)$$

Stationarity of L''_B with respect to Lagrange multiplier μ yields the normality condition (4). On substituting λ from (30) into (4) we obtain

$$\exp(\rho\mu - 1) \sum_{j=1}^M \exp(\rho\alpha g_j(\mathbf{x})) = 1 \quad (31)$$

The term $\exp(\rho\mu - 1)$ may be eliminated between (30) and (31) to give

$$\lambda_k = \exp(\rho\alpha g_k(\mathbf{x})) / \sum_{j=1}^M \exp(\rho\alpha g_j(\mathbf{x})) \quad k = 1, \dots, M \quad (32)$$

It can be seen that (32) is identical to the λ -update formula (19) used in the two-phase surrogate algorithm if $\beta = \rho\alpha$. In the earlier algorithm we required β to be an increasing positive number.

Stationarity of L''_B with respect to α gives the surrogate constraint (3) which, together with (7), remain to be solved. Substituting λ from (32) into (3) and (7) gives $N + 1$ equations in the $N + 1$ variables \mathbf{x} and α . Though these may in theory be solved they are non-linear and intractable so, instead, we return to the Lagrangean L''_B itself. Substituting λ from (32) into (28) gives

$$L''_B = f(\mathbf{x}) + \alpha \left\{ \sum_{j=1}^M g_j(\mathbf{x}) \exp[\rho\alpha g_j(\mathbf{x})] \right\} / \sum_{j=1}^M \exp[\rho\alpha g_j(\mathbf{x})] - \frac{1}{\rho} \frac{\sum_{j=1}^M (\exp[\rho\alpha g_j(\mathbf{x})] \ln\{\exp[\rho\alpha g_j(\mathbf{x})] / \sum_{j=1}^M \exp[\rho\alpha g_j(\mathbf{x})]\})}{\sum_{j=1}^M \exp[\rho\alpha g_j(\mathbf{x})]}$$

After much algebraic simplification this reduces to

$$L''_B = f(\mathbf{x}) + \frac{1}{\rho} \ln \sum_{j=1}^M \exp[\rho\alpha g_j(\mathbf{x})] \quad (33)$$

We have already noted that ρ in (33) is an arbitrary positive constant associated with the entropy augmentation term in Eq. (28) and that $\rho\alpha$ should be an increasing positive quantity. Thus for fixed positive ρ , α should be increasing positive. This is in the spirit also of the duality relationship (8), and so leads to a single-phase formulation as:

$$\left. \begin{array}{l} \text{Minimize}_{\mathbf{x}, i=1, \dots, N} \quad f(\mathbf{x}) + \frac{1}{\rho} \ln \sum_{j=1}^M \exp[\rho\alpha g_j(\mathbf{x})] \\ \text{with an increasing, positive } \alpha. \end{array} \right\} \quad (34)$$

8 COMMENTS

Problem (34) has not yet been studied in depth in respect of its algorithmic performance. Nevertheless, some comments on it are warranted. The unconstrained function is a curious hybrid between augmented Lagrangian, penalty function, and exact (single iteration) penalty function and several different algorithmic treatments of it seem possible. Choosing a constant value for ρ and minimizing over \mathbf{x} with occasional updates of the increasing positive α is one possibility, as indicated in (34). It also seems possible to treat (34) as a straight-forward multi-iteration penalty function, choosing a fixed α and minimizing over \mathbf{x} with a sequence of increasing positive ρ .

The "penalty" term in (34) is worth examining. For high positive $\rho\alpha$ it clearly penalises constraint violations ($g_j(\mathbf{x}) > 0$), whilst inactive constraints ($g_j(\mathbf{x}) < 0$) contribute negligibly small amounts to the total penalty. The $\ln \sum \exp$ form of the penalty term is particularly interesting as, to the authors' knowledge, it has not appeared before among the many suggested penalty functions in the mathematical programming literature. However, in 1979, Kreisselmeier and Steinhauser¹⁶ introduced this precise form

$$\frac{1}{\rho} \ln \sum_{j=1}^M \exp[\rho\alpha g_j(\mathbf{x})] \quad (35)$$

without any indication of its derivation or origins except that it displayed some desired characteristics, in a multi-criteria optimization context. The form (35) has subsequently been used to considerable effect as a type of cumulative constraint by Sobieski¹⁷ and Hajela,^{18,19} citing Kreisselmeier and Steinhauser¹⁶ as its source. It is particularly interesting that, unknown to those who have used it, the Kreisselmeier-Steinhauser function (35) turns out to be so closely linked by virtue of this present work to entropy maximization processes. Further papers by the present authors will address these links.

9 CONCLUSIONS

This paper has explored the possibility that information theory and informational entropy maximization processes may have a place in the development of new algorithms for solving mathematical programming problems. In this introductory paper entropy maximization was incorporated into non-linear constrained optimization through a surrogate constraint approach. An information-theoretic interpretation and justification for this was presented which sheds some new light on old methods.

Extensions of the surrogate entropy method were presented which resulted in some interesting links with other recent developments in engineering optimization. Much further work remains to be done, but this paper has shown that informational entropy processes and mathematical programming are closely linked and that further study is warranted. The need for some radically different and improved optimization techniques is now urgent in the area of large-scale engineering design. The present work represents one new direction which might possibly result in improved techniques in the future.

REFERENCES

1. H. Everett, "Generalized Lagrange multiplier method for solving problems of optimum allocation of resources." *Opns. Res.*, **11**, (2), 399-417, (1963).
2. R. Brooks and A. Geoffrion, "Finding Everett's Lagrange multipliers by linear programming." *Opns. Res.*, **14**, (6), 1149-1153, (1966).
3. F. J. Gould, "Extensions of Lagrange multipliers in nonlinear programming." *SIAM J. Appl. Maths*, **13**, (6), 1280-1297, (1969).
4. H. J. Greenberg and W. P. Pierskalla, "Surrogate mathematical programming." *Opns. Res.*, **18**, (5), 924-939, (1970).
5. E. T. Jaynes, "Information theory and statistical mechanics." *The Physical Review*, **106**, 620-630 and **108**, 171-190, (1957).
6. P. W. Jowitt, "Bayesian estimates of material properties from limited test data." *Eng. Structures*, **1**, (4), 170-178, (1979).
7. P. C. Basu and A. B. Templeman, "Structural reliability and its sensitivity." *Civ. Eng. Systems*, **2**, (1), 3-11, (1985).
8. J. Munro and P. W. Jowitt, "Decision analysis in the ready-mixed concrete industry." *Proc. Inst. Civ. Engrs., Part 2*, **65**, 41-52, (1978).
9. S. Guiasu, "Maximum entropy condition in queuing theory." *J. Opnal. Res. Soc.*, **37**, (3), 293-301, (1986).
10. M. Tribus, "Information theory as the basis for thermostatics and thermodynamics." *ASME, J. Appl. Mech.*, 1-8, (1961).
11. C. E. Shannon, "A mathematical theory of communication." *Bell System Technical Journal*, **27**, (3), 379-428, (1948).
12. A. B. Templeman and X. S. Li, "Entropy duals." *Engineering Optimization*, **9**, (2), 107-119, (1985).
13. C. S. Beightler and D. T. Phillips, *Applied Geometric Programming*, John Wiley, 239-242, (1976).
14. F. Glover, "Surrogate constraints." *Opns. Res.*, **16**, (4), 741-749, (1968).
15. F. Glover, "Surrogate constraint duality in mathematical programming." *Opns. Res.*, **23**, (3), 434-451, (1975).
16. G. Kreisselmeier and R. Steinhauser, "Systematic control design by optimizing a vector performance index." *IFAC Symp. Computer Aided Design of Control Systems*, Zurich, Switzerland, August 29-31, (1979).
17. J. Sobieszczanski-Sobieski, B. B. James and A. R. Dovi, "Structural optimization by multilevel decomposition." *AIAA J.*, **23**, (11), 1775-1782, (1985).
18. P. Hajela, "Further developments in the controlled growth approach for optimal structural synthesis." *ASME, paper 82-DET-62*, (1982).
19. P. Hajela, "Techniques in optimum structural synthesis with static and dynamic constraints." *Ph.D. Thesis*, Stanford University, (1982).