AMC PAMPHLET

AMCP 706-192

ENGINEERING DESIGN HANDBOOK

COMPUTER AIDED DESIGN OF MECHANICAL SYSTEMS

HEADQUARTERS, U.S. ARMY MATERIEL COMMAND

JULY 1973

HEADQUARTERS
US ARMY MATERIEL COMMAND
5001 EISENHOWER AVENUE
ALEXANDRIA, VA 22304

AMC PAMPHLET
NO. 706-192

15 July 1973

ENGINEERING DESIGN HANDBOOK
COMPUTER AIDED DESIGN
OF MECHANICAL SYSTEMS

TABLE OF CONTENTS

TABLE OF CONTENTS (Con't.)

Paragraph *Page*

CHAPTER 2. FINITE DIMENSIONAL UNCONSTRAINED OPTIMIZATION

TABLE OF CONTENTS (Con't.)

TABLE OF CONTENTS (Con't.)

TABLE OF CONTENTS (Con't.)

TABLE OF CONTENTS (Con't.)

TABLE OF CONTENTS (Con't.)

TABLE OF CONTENTS (Con't.)

LIST OF ILLUSTRATIONS

LIST OF ILLUSTRATIONS (Con't.)

LIST OF ILLUSTRATIONS (Con't.)

LIST OF ILLUSTRATIONS (Con't.)

LIST OF TABLES

LIST OF TABLES (Con't.)

PREFACE

The Engineering Design Handbooks of the U. S. Army Materiel Command are a coördinated series of handbooks containing basic information and fundamental data useful in the design and development of Army materiel and systems.

This text treats a broad class of optimal design problems through use of a consistent set of computational techniques ideally suited for computer application to mechanical design problems. No attempt has been made to be exhaustive in the treatment of optimization techniques or the full range of mechanical applications. Rather, the class of problems treated is concisely formulated (in Chapters 4 and following) in terms of design and state variables that occur in mechanical design. A steepest-descent approach – which has served as a workhorse, reliable technique in fields such as aerodynamic system design, control theory, and nonlinear programming – is developed here for mechanical system design.

Extensive application of design optimization techniques is made in the field of structural design, as well as in a limited number of specific weapon design problems. The examples are presented in considerable detail, as they are encountered in practice, to provide the practicing engineer with insight into use of the methods for his class of problems. A consistent design philosophy is maintained throughout the text to assist the designer in extrapolating the methods to classes of problems that are only similar mathematically to the examples treated here.

The text is structured so that it can be understood and used by practicing engineers with a good background in calculus and matrix theory. Computational algorithms are stated in considerable detail so that they can be effectively implemented by junior engineers, with only problem formulation and general supervision provided by a senior project engineer. As with virtually all computer aided design techniques, some computing art is required for effective implementation of these techniques. The detailed treatment of structural applications in Chapters 5, ', and 9 should provide insight into this computational art. References are given to more advanced literature for proofs of theorems and extensions of methods to other classes of problems.

The Handbook was written by Dr. Edward J. Haug, Jr. of the U. S. Army Weapons Command. It is based on lecture materials used by him in a two-semester graduate sequence on "Optimization of Structural Systems", taught at the University of Iowa since 1968. Examples treated in the text are derived primarily from Dr. Haug's research, Dr. Jasbir Arora's University of Iowa dissertation, and the work of Messrs. Tom Streeter and Stephen Newell of the U.S. Army Weapons Command.

The Engineering Design Handbooks fall into two basic categories, those approved for release and sale, and those classified for security reasons. The Army Materiel Command policy is to release these Engineering Design Handbooks to other DOD activities and their contractors and other Government agencies in accordance with current Army Regulation 70-31, dated 9 September 1966. It will be noted that the majority of these Handbooks can be obtained from the National Technical Information Service (NTIS). Procedures for acquiring these Handbooks follow:

a. Activities within AMC, DOD agencies and Government agencies other than DOD having need for the Handbooks should direct their request on an official form to:

Commander
Letterkenny Army Depot
ATTN: AMXLE-ATD
Chambersburg, PA 17201

b. Contractors and universities must forward their requests to:

National Technical Information Service
Department of Commerce
Springfield, VA 22151

(Requests for classified documents must be sent, with appropriate "Need to Know" justification, to Letterkenny Army Depot.)

Comments and suggestions on this Handbook are welcome and should be addressed to:

US Army Materiel Command
ATTN: AMCRD-TV
5001 Eisenhower Avenue
Alexandria, VA 22304

DA Forms 2028 (Recommended Changes to Publications), which are available through normal publications supply channels, may be used for comments/suggestions.

CHAPTER 1

ELEMENTS OF COMPUTER AIDED DESIGN

1-1 SYNTHESIS VS ANALYSIS IN ENGINEERING DESIGN

Engineering is defined (Ref. 1) as "the art or science of making practical application of the knowledge of pure sciences such as physics, chemistry, biology, etc.". Although broad, this definition implies that the job of engineering is to synthesize, or put together, useful systems by applying knowledge and methods derived from the "pure" sciences. The meaning of "practical" in the given definition should be interpreted as best, or optimal; i.e., the job of engineering design is to develop the best possible system for the given application, consistent with the resources allocated to the development phase. The purpose of this handbook is to present a class of methods that allow for efficient use of the computer in the design process.

Since the computer can be viewed simply as a device to handle large quantities of data and perform simple algebraic operations and logic rapidly, it is important to look first into the role of calculation in design. The usual approach to design is to conceive of a candidate system and then test it to see if it works. Great strides have been made with digital computers in the past two decades to allow for numerical analysis as a test of the idea, or concept, rather than previous cut-and-try techniques. For example, in structural design one chooses the configuration and member sizes, and then tests the structure by analyzing its response to given loads. If the structure does not behave as desired, then design changes are made and the structure is reanalyzed. This process continues until the designer is satisfied with his design. This has been the principal use of the computer in the design process.

In general, then, before the designer can assure himself that he has the best system, he must be capable of analyzing all candidates. In the past half century, outstanding advances in engineering analysis have been made. The digital computer has allowed the engineer to quantitatively analyze the behavior of systems that were examined only qualitatively in the past. The mechanical sciences, particularly, have benefited from this boom in analysis capability. Structural analysis, stress analysis, analysis of mechanisms, and heat transfer analysis, just to name a few, have made spectacular advances in the past twenty years.

Until the early 1960's, and even to the present day to a lesser extent, the attention of engineering research has been focused primarily on developing analysis capability. During this period of emphasis on analysis, inadequate attention was paid to developing a synthesis, or design, capability that is able to efficiently use the newly developed analysis methods. In some of the mechanical sciences, this problem is particularly acute. In structural mechanics, for example, it is possible to analyze a structure under a given loading to obtain accurate values for stress, displacement, and even natural frequency. It is not clear, however, how a structure should be laid out and proportioned to efficiently utilize material in order to meet strength requirements. A more difficult problem is the proportioning of a structure so as to efficiently limit displacement and meet constraints on natural frequency and buckling. For a review of the state of optimal structural design through 1967, see Ref. 2.

It appears that the analysis capability needed for computer aided design is available. The next problem to be addressed, then, is the matter of what is meant by best, or optimum. The idea of best enters very naturally into engineering design efforts. In profit-motivated industries as well as in Government laboratories, the objective is to maximize some return function while satisfying constraints such as resource allocation, performance requirements, and human limitations.

Once some return function or measure of value is chosen and constraints are identified, the system designer would like to have some optimal design methodology that is capable of aiding him in the determination of the best, or practically best, system. It must be emphasized at this point that the search is not for an automatic optimization technique that can solve any design problem fed to it. Rather, the need is for an optimal design methodology that can aid the engineer in the implementation of his concepts and guide him in a direction which, if continued indefinitely, would yield a mathematical optimum.

A key challenge to developers of practical computer aids to designers is to take maximum advantage of human judgment in the design process. The potential of interactive computation and design information display is only now in a developing stage and holds promise for significant improvement of the value of the computer in design.

1-2 THE PHILOSOPHY OF SYSTEM ENGINEERING

In the middle 1950's a formalized approach to the development of large-scale, man-made systems began to appear in the literature, see Refs. 3, 4, 5. This approach, which has features common to most problem solving processes, was given the name "system engineering" and is the very essence of computer aided design. A feature which sets system engineering and computer aided design off from most of the logical problem solving schemes is the explicit inclusion of key considerations peculiar to engineering design of systems. A second important feature of system engineering is the attention paid to quantitative description of the system and its behavior.

The basic idea in system engineering is to begin with a statement of system requirements and objectives, and move in an organized way toward an optimum system. A process which illustrates the approach is shown in Fig. 1-1.

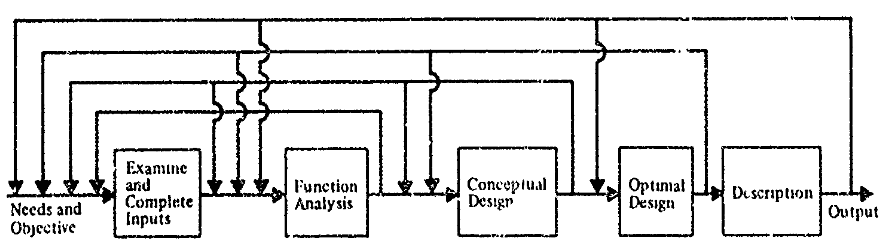


Figure 1-1. A System Engineering Model

The purpose of this text is not to give a detailed treatment of system engineering, but rather to present aspects of the theory of computer aided design, with emphasis on optimal design. The simplified model of a system engineering process shows that optimal design is a part of system engineering, but, indeed, by no means the dominant part. The purpose of this paragraph is to discuss the interface of optimal design with the remaining essential elements of system engineering.

System engineering begins with the identification of a need by a potential user of the system to be developed. It is often the case that the user knows that he needs a system to do a job, but he may have difficulty in stating his needs and objectives quantitatively. It then becomes the joint responsibility of the system engineer and user to quantify system objectives so that a meaningful set of objectives may be established for the development to follow.

Once the needs and objectives for a system are identified, it is necessary to define functions that must be performed by the system and any subsystems that are required. This process is called function analysis, and its purpose is to pick out functions or operations that must be performed in order to accomplish the mission required of the system being developed. These functions then become lower level objectives for the development of subsystems. Identification of functions tends to be qualitative in nature. However, once a function or operation is identified, it must be described in quantitative terms, if at all possible. For example, if a function must occur quickly, the time allowed should be specified.

The next step shown in Fig. 1-1 is one of the most difficult functions in system engineering and certainly the most difficult step to describe analytically. Conceptual design, as its name implies, is the identification of the various concepts or basic system configurations that might meet the system objectives. It is desirable in this step to leave the concepts as general as possible so as not to eliminate candidate systems that might be very effective. For example, if the function to be performed is to propel a vehicle over the surface of the earth, conceptual designs might include wheels, tracks, legs, air cushion, etc.

It is important at this time to identify ranges of values of parameters describing the system so that, for any parameter in this range of values, the system will perform the functions identified in the previous step, i.e., the set of parameters describing admissible systems is identified. It is at this time that the experienced designer can be extremely valuable in reflecting state-of-the-art capabilities of technologies involved in the system development.

The optimal design step has as its objective the choice of the undetermined parameters identified in the previous step. These parameters must be in the ranges defined by technological limitations and system functions. The criterion for choosing system parameters is maximization of system worth or value. It should be emphasized that a mathematically precise optimum may be impossible to attain and must therefore serve only as a goal. Methods for choosing system parameters should, however, have the property that if an optimum does exist, then given enough patience and computer time, that optimum should be approached as a limit.

What appears to be the final step in the system engineering model of Fig. 1-1 Description, is, in reality, probably just an intermediate step. Unless the system design procedure has been unusually effective, the system decided upon will probably not satisfy the user. More likely, it will probably not satisfy the system engineering team. Having the results of one pass through the system engineering process, the user can probably remember some constraints which he forgot to specify and which the optimum system violates. The designer probably also will see concepts that he did not see before. Much as the user, he too will remember technological constraints which he forgot to specify and which the optimum system violates. Finally, the sponsoring activity will undoubtedly decide that it will be all right to decrease the measure of system value a small amount if it will save some money.

The next step in the procedure is for each member of the team to take a deep breath, sigh, and go back to work, armed with his hard earned new knowledge. It is for this purpose that all the feedback paths in the model of Fig. 1-1 are shown. This iterative procedure is then continued until the sponsoring activity decides that the system developed is what it really needs. This will probably be another human decision, rather than a programmed mathematical one.

The remaining chapters will be devoted to the problem of computer aided and optimal design. If the design methods presented later are to be of maximum value to the reader, he must have a clear picture of how these methods fit into the larger problem of system engineering. For further literature on the basic ideas involved in system engineering, see Refs. 3, 4, and 5.

1-3 COMPUTER AIDED DESIGN IN THE MECHANICAL SCIENCES

The theory of computer aided and optimal design is developed in subsequent chapters as it applies to the mechanical sciences. There are peculiarities of mechanical design, as opposed to classical control system design, which require specialized treatment. Further, the mathematics involved in mechanical system design is quite different from the mathematics of control theory. These distinctions are highlighted throughout the text.

In the chapters that follow, optimal control theory is interpreted as treating feedback controllers; i.e., an optimal control system has active elements that sense errors in output, due to fluctuations in inputs, and adjust system controls so as to maximize some measure of system performance. Optimal design, on the other hand, is taken as the problem of choosing system elements or parameters describing these elements, which are fixed for the life of the elements, so that the system is optimum in some sense. In control literature this is called open loop control. The principal difference in the two problems is that the variables chosen in the optimal design problem are fixed for the life of the system, whereas variables in a feedback control device are to be adjusted according to inputs as the system operates. Mathematically, the difference in the two results is that the control law describes how the system variables should be adjusted as a function of the state of the system, whereas an optimum design is simply a set of parameters describing system elements and will not be changed during the life of the system. This distinction is not uniform in the control literature but is used here to identify the class of problems treated.

In most literature on control problems, sequential systems are treated, i.e., operations of the system progress one after another as if they were occurring in time in a pre-arranged order. Many optimal design problems are not of this kind. For example, in designing a structure one must be concerned with stresses due to applied loads. These stresses are interpreted as the state of the structural system. They are determined by a boundary-value problem that cannot be interpreted as a sequential process (initial-value problem). In some design problems it is possible to define auxiliary variables so that the governing equations form an initial-value problem with additional constraints. This procedure, however, generally complicates the problem unnecessarily. For this reason the problems in succeeding chapters are formulated as boundary as opposed to initial-value problems.

In order to illustrate the use of the methods presented, applications are made primarily in optimal structural design. Applications are chosen to illustrate the use of the methods on problems having a number of design variables which might be found in engineering applications. Further, since many of the methods are relatively new, it is anticipated that improvements in computational efficiency may be realized in specific problems if advantage is taken of special features of the class of problems treated.

It is appropriate to highlight a significant computational distinction between two classes of design problem. The reader may note that Chapters 2 through 5 of this text deal with problems in which system design and performance are specified by a finite number of parameters (real numbers). Chapters 6 through 9, on the other hand, deal with systems that are described by functions on some given space or time domain. Mathematically, these problems are called finite and infinite dimensional, respectively. Optimization theory for these two classes of problems can be put in the same form, but there are very real differences in the computational techniques available for design optimization. Since the subject of this handbook is computer aids to design, the practical distinction is made here. For a unifying mathematical treatment, the reader is referred to Ref. 7.

Finally, it is important to realize that engineering design optimization and engineering analysis are fundamentally different in nature. In analysis, one is generally assured that a solution exists and numerical methods are generally stable. In optimal design, on the other hand, existence of even a nominal design satisfying objectives is not assured, much less existence of an optimal design. Moreover, even when an optimum exists, numerical methods for its solution are often quite sensitive to initial estimates and require much computational art for iterative convergence. These properties will be observed over and over in this handbook when example problems are treated.

It is important that the reader take a mathematical outlook when doing computer aided design and optimization. A purely intuitive approach can lead to erroneous results that may not be apparent until someone happens onto a nominal design which is vastly superior to a "supposed" optimum design.

1-4 MATHEMATICAL PRELIMINARIES

The level of mathematical background required for an understanding of the methods of optimal design presented in the following chapters is a course in advanced calculus and the ability to use matrix notation. Since engineers often require results of rather deep mathematical analyses to solve real-world problems, several results have been accepted with references given to proofs. The purpose of this paragraph is to present notation and some basic mathematical ideas used throughout the text.

Since most real-world problems involve several parameters, it is convenient to utilize vector notation. For example, rather than writing $x_1, ..., x_n$ repeatedly, these n variables are collected into a column vector

$$x = \begin{bmatrix} x_1 \\ \cdot \\ \cdot \\ x_n \end{bmatrix} . \tag{1-1}$$

Unless otherwise noted, all vector variables will be column vectors. A vector of the form (Eq. 1-1) may be interpreted as a point in n-dimensional real space, R^n. The space R^n is simply the collection of all n-vectors of real numbers. For example, the real line is R^1 and the plane is R^2.

It will often be convenient to deal with a collection of points in the space R^n. A collection of points D in R^n will be called a set, or a subset of R^n. A point x in R^n which is in D will be denoted $x \in D$. This will be the extent of set notation required in later chapters.

In R^n there is a well defined idea of length of a vector. This analog of length in the real world will be denoted

$$\| x \| \equiv \left[\sum_{i=1}^{n} x_i^2 \right]^{1/2} \tag{1-2}$$

and is called a norm on R^n. There are many norms defined on R^n but Eq. 1-2 will be sufficient for the purposes of this text. Along with the idea of norm on R^n goes the concept of dot product or inner product. The inner product of two elements x and y of R^n is

$$< x, y > \equiv x^T y = \sum_{i=1}^{n} x_i y_i \quad . \tag{1-3}$$

Two vectors are called orthogonal if their inner product is zero.

The idea of convergence of a sequence $\{ x^i \}$ in R^n with norm (Eq. 1-2) is much like convergence of real numbers. That is, $\lim_{i \to \infty} x^i = x$ if for any $\epsilon > 0$ there is an $N > 0$ such that $\| x^i - x \| < \epsilon$ for all $i > N$. An important property of sets in optimization theory is closedness. A subset D of R^n is called closed if every sequence in D which converges has its limit in D.

Just as the idea of collecting n real numbers into a vector in R^n, it is helpful to define a vector function $g(x)$ for $x \in R^n$ as

$$g(x) = \begin{bmatrix} g_1(x) \\ \cdot \\ \cdot \\ g_m(x) \end{bmatrix} . \tag{1-4}$$

Such a function is called continuous at $\bar{x}$ if for any $\epsilon > 0$ there is a $\delta > 0$ such that $\| g(x) - g(\bar{x}) \|_m < \epsilon$ if $\| x - \bar{x} \|_n < \delta$. The subscripts m and n on the norms denote the dimension of the space on which the norm is defined.

It will often be desirable to deal with a set of functions which satisfy

$$g_i(x) \leqslant 0, i = 1, ..., m \quad . \tag{1-5}$$

In this case it is convenient to define inequality among vectors as

$$g(x) \leqslant 0 \tag{1-6}$$

where inequality is taken componentwise, i.e., Eq. 1-6 is defined to mean the same thing as Eq. 1-5.

One of the most useful notations in the following chapters is the idea of the derivative of a vector function with respect to its vector variable. This notation is

$$\frac{dg(x)}{dx} \equiv \left[\frac{\partial g_i(x)}{\partial x_j}\right] m \times n \tag{1-7}$$

where i is a row index and j is a column index. If $f(x)$ is a real valued function of $x \in R^n$, this notation is

$$\frac{df(x)}{dx} \equiv \left[\frac{\partial f(x)}{\partial x_1}, \ldots, \frac{\partial f(x)}{\partial x_n}\right]. \tag{1-8}$$

The derivative of a real valued function is often called the gradient of that function and is denoted

$$\nabla f(x) \equiv \frac{df(x)}{dx}. \tag{1-9}$$

The gradient is one of the few standard symbols which denotes a row vector rather than a column vector. Likewise, for a real valued function the matrix of second derivatives may be defined as the matrix

$$\frac{d^2 f(x)}{dx^2} \equiv \nabla^2 f(x) \equiv \left[\frac{\partial^2 f(x)}{\partial x_i \partial x_j}\right]. \tag{1-10}$$

An important theorem in the analysis of functions appearing in optimal design problems is Taylor's Theorem.

Taylor's Theorem: Let the real valued function $f(x)$ have $k+1$ continuous derivatives in R^n. Then for $x, y \in R^n$, there is a point $\xi = \alpha x + (1-\alpha)y$ with $0 < \alpha < 1$, such that

$$f(y) = f(x) + \sum_{i=1}^{n} \frac{\partial f(x)}{\partial x_i}(y_i - x_i) \tag{1-11}$$

$$+ \frac{1}{2}\sum_{j=1}^{n}\sum_{i=1}^{n} \frac{\partial^2 f(x)}{\partial x_i \partial x_j}(y_i - x_i)(y_j - x_j)$$

$$+ \ldots + \frac{1}{k!} \sum_{i_1+\ldots+i_n=k} \frac{\partial^k f(x)}{\partial x_1^{i_1} \ldots \partial x_n^{i_n}} \times$$

$$(y_1 - x_1)^{i_1} \ldots (y_n - x_n)^{i_n}$$

$$+ \frac{1}{(k+1)!} \sum_{j_1+\ldots+j_n=k+1} \frac{\partial^{k+1} f(\xi)}{\partial x_1^{j_1} \ldots \partial x_n^{j_n}} \times$$

$$(y_1 - x_1)^{j_1} \ldots (y_n - x_n)^{j_n}.$$

For proof of this theorem see Ref. 6, page 56.

In many places in the following chapters, Taylor's Theorem will be used to obtain an approximate expression for a function at a point sufficiently near a point where the function is known. The most common approximation is the one obtained by deleting second and higher order terms. For example, if $\|x - y\|$ is small,

$$f(y) - f(x) \simeq \frac{df(x)}{dx}(y - x) \tag{1-12}$$

where by Eq. 1-11 the error in Eq. 1-12 is at most a constant times $\|y - x\|^2$ if $f(x)$ has bounded second order derivatives. The left side of Eq. 1-12 is generally denoted by $\delta f(x)$, where $y - x$ is denoted δx. In this notation,

$$\delta f(x) = \frac{df}{dx}\delta x. \tag{1-13}$$

Eq. 1-13 may be thought of as a total differential. Even for vector functions $g(x)$, Eq. 1-13 holds for each component so if

$\delta g(x) \equiv [\delta g_1(x), \ldots, \delta g_m(x)]^T$, then

$$\delta g(x) = \frac{dg(x)}{dx}\,\delta x. \tag{1-14}$$

In later work, $g(x)$ will often be a function of $x \in R^n$ and $z \in R^p$. In this case, Eq. 1-14 is

$$\delta g(x,z) = \frac{\partial g(x,z)}{\partial x}\,\delta x + \frac{\partial g(x,z)}{\partial z}\,\delta z \tag{1-15}$$

where

$$\frac{\partial g(x,z)}{\partial x} = \left[\frac{\partial g_i(x,z)}{\partial x_j}\right]_{m \times n}$$

and

$$\frac{\partial g(x,z)}{\partial z} = \left[\frac{\partial g_i(x,z)}{\partial z_j}\right]_{m \times p}$$

Most of the common notation used in later chapters now has been defined. Special notation and results required locally for some development will be defined and used there.

1-5 ILLUSTRATIVE MILITARY COMPUTER AIDED DESIGN PROBLEMS

In this paragraph two illustrative military optimal design problems are formulated, and computer aided design techniques are outlined for their solution. The treatment here is only for the purpose of introducing concepts. These examples are treated in more depth in Chapters 7 and 8.

1-5.1 OPTIMAL DESIGN OF STRUCTURES

The optimization technique described in this paragraph was initially developed for application to minimum weight structural design problems. For this reason, and to give an engineering feel for application of the technique, the method will be presented along with examples from the field of optimal structural design.

As a specific example, let us consider a design problem whereby a highly directional transmission device, or perhaps a gun, is to be mounted on a tower or gun mount that is required to support the device at some given distance away from the basic supporting structure, such as the earth. A schematic of the problem is shown in Fig. 1-2. The basic

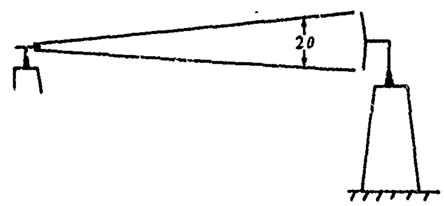


Figure 1-2. Structural Requirement

problem is to design a structure that supports the device under consideration and which is as light as possible for purposes of transportation and erection on the battlefield, or perhaps mounting on a helicopter. A basic design requirement for this structure is that the device mounted on the top shall not have an angular deflection of more than θ radians, in order to hit the receiver or target. The loading that is to be considered is a wind load of up to a given velocity, which would cause angular deflection of the top of the tower.

The needs and objectives in this design problem are well established, so no additional inputs need be considered at the present time. Further, the requirement that the tower support the device with only a given allowable angular deflection is the only basic function required of the tower; thus the function analysis block of Fig. 1-1 is also complete. The next stage, and one that is quite difficult to describe analytically, is that of arriving at conceptual towers which might perform the given mission.

Four different conceptual designs are shown in Fig. 1-3. The first two concepts, Figs. 1-3(A) and (B), involve rigidly fixing the tower at its base to the fundamental supporting structure. In both towers, variable spacing as a function of height is allowed between vertical members of the structure. In addition,

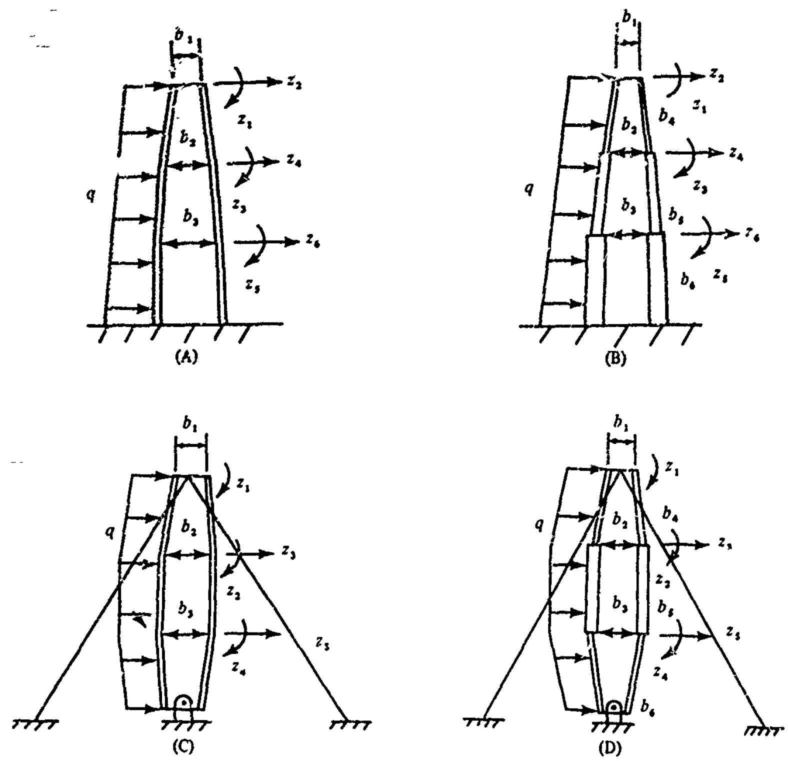


Figure 1-3. Conceptual Designs

one of the concepts allows for varying the area of the main structural members as a function of height. The second set of concepts, Figs. 1-3(C) and (D), involves towers that are pinned at their base to the supporting structure and that are supported by guy wires at the top of the structure. Likewise, in both of these concepts, variable spacing of the main vertical members is allowed. In the second concept, variation of area along the length of the tower is also allowed. It should be noted that the conceptual designs in Fig. 1-3 can have as many subsections with different area and spacing as desired. Three are shown for convenience in the figure.

In each of the conceptual towers of Fig. 1-3, the variables b_1 through b_3 describe the variable spacing of the members of the tower. In two of the concepts, Figs. 1-3(B) and (D), b_4 through b_6 specify the variable areas in the construction of the main vertical member. These variables serve as design parameters, in that the designer can choose these variables and completely specify the design of the tower.

In addition to the design variables, a main part of the design problem is the behavior of the structure under wind load, since one of the major constraints on behavior of the structure is that the angular deflection of the top of the tower not exceed an angle θ. For this reason, the angular deflection of each of the joints must be determined, along with lateral deflection due to lateral wind loading. This is a relatively routine analysis problem when one uses the techniques of finite element structural analysis. Not shown in Fig. 1-3, but required in the construction, are cross members which maintain spacing of the main vertical members. In order to state the optimal design problem mathematically, first define vectors of design variables b_i and state variables z_i

$$\begin{aligned} b &= [b_1, b_2, \ldots b_m]^T \\ z &= [z_1, z_2, \ldots z_n]^T . \end{aligned} \tag{1-16}$$

Using finite element structural analysis techniques, define the stiffness matrix as

$$A(b) = [a_{ij}(b)]_{n \times n} \tag{1-17}$$

where the dependence of stiffness on the design variables is explicitly shown. Using this matrix, the structural response is given by the following matrix equation

$$A(b)z = q \tag{1-18}$$

where q is the wind loading matrix.

Now that the relationship between the design variables and the structural response is specified by Eq. 1-18, the next step in formulating an optimal design problem is the identification of constraints. In order to prevent dimensions or structural areas from going to zero, resulting in an unstable structure, it is required that the design variables be bounded uniformly away from zero. This is given formally by the inequality

$$b_i > b_{io} > 0, i = 1, \ldots, m. \tag{1-19}$$

The fundamental constraint in the present problem is that the angular deflection at the top of the tower not exceed the angle θ. This is expressed analytically by the inequality

$$|z_1| < \theta. \tag{1-20}$$

The final step in formulation of an optimal design problem is to identify the cost function to be minimized. In the present case, the cost function is structural weight J and is given by an expression of the form

$$J = \gamma \sum_{i=1}^{m} c_i b_i \tag{1-21}$$

where γ is material density and c_i are weighting factors representing lengths of structural elements and weight requirements for lateral stiffners.

We now have an optimal structural design problem that is well formulated from a mathematical point of view. The objective is to find the design variables b_1 through b_m that satisfy constraint Eqs. 1-19 and 1-20, and which minimize the structural weight as given by Eq. 1-21. The technique used to solve this problem, and in fact a large class of optimal system design problems, is based on a very simple idea of engineering design. The idea of the technique is to allow small variations in some nominal design, and analyze the effect of these variations on the equations of the problem and the cost function associated with the problem. As a result of allowing only small design changes, certain approximations may be made that allow the best change in design variables to be determined in order to decrease the cost function of the problem as much as possible, while still not violating constraints of the design problem. For example, one might choose as an initial estimate of the optimal design a uniform tower as shown in Fig. 1-4. The estimated design variable in this case is denoted by $b^{(0)}$.

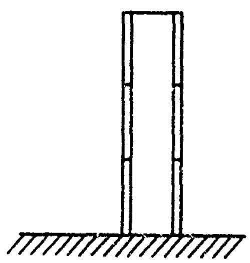


Figure 1-4. Uniform Initial Design

Let δb be a small change in the design variable $b^{(0)}$. Any change in the design variable will result in a change in the structural response, denoted by δz. The nature of the structural analysis problem guarantees that small δb yields small δz. Further, a Taylor series approximation of terms appearing in Eq. 1-18 yields

$$A(b^{(0)})\delta z + \frac{\partial}{\partial b}(A(b)z|_{b=b^{(0)}})\delta b = 0. \quad (1\text{-}22)$$

If an inequality constraint is violated, such as

$$b_i < b_{io} \quad (1\text{-}23)$$

then in order to correct the constraint error it is required that

$$\delta b_i > b_{io} - b_i. \quad (1\text{-}24)$$

Or, if the angular deflection constraint is violated, for example,

$$z_1 > \theta \quad (1\text{-}25)$$

then, to correct the constraint error it is required that

$$\delta z_1 < \theta - z_1. \quad (1\text{-}26)$$

Finally, the change in structural weight due to the change in design δb is given by

$$\delta J = \gamma \sum_{i=1}^{m} c_i \delta b_i. \quad (1\text{-}27)$$

The object of the new problem is to determine δb so as to minimize the linearized cost function of Eq. 1-27, subject to constraint Eqs. 1-24 and 1-26. Due to the special nature of this problem, the optimum change δb can be determined in closed form. For a detailed derivation of this optimum perturbation, the reader is referred to Chapter 5. For discussion here, the results of this calculation will be denoted by

$$\delta b = \eta B + C \quad (1\text{-}28)$$

where the vectors B and C depend on $b^{(0)}$, constraint errors, and equations of the problem. The parameter η is an undetermined parameter that plays the role of a step size, when viewed in the geometry of design variable space. For example, if there are only two design parameters b_1 and b_2, the direction of the desired change is shown by B in Fig. 1-5, and η is a step size along that direction. In the terminology of optimization theory, B is known as the direction of steepest descent. It is analogous to the direction one would go downhill in order to reduce

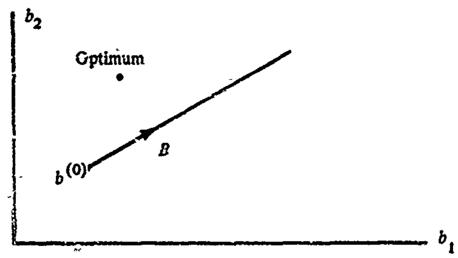


Figure 1-5. Direction of Steepest Descent

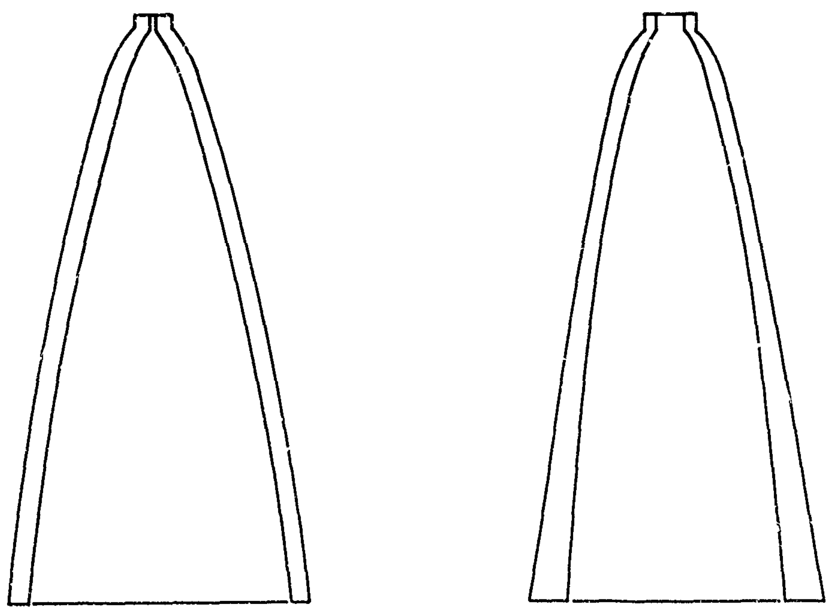
his altitude as rapidly as possible. It is clear that on normal hills, as in most design cost functions, the direction of steepest descent changes, depending on the location on that hill. For this reason, the direction of steepest-descent does not generally pass through the optimum point as shown in Fig. 1-5.

There are many techniques for choosing the step size η. The one used in the steepest descent method is based on requesting a certain reduction in the cost function due to the changed δb. This request, then, determines the step size η and one can calculate δb from Eq. 1-28. This δb is the best change in the estimated design variable $b^{(0)}$. This best change is then added to the initial estimate to obtain a new estimate that corresponds to a structure of less weight and that still satisfies the constraints of the problem, i.e.,

$$b^{(1)} = b^{(0)} + \delta b \qquad (1\text{-}29)$$

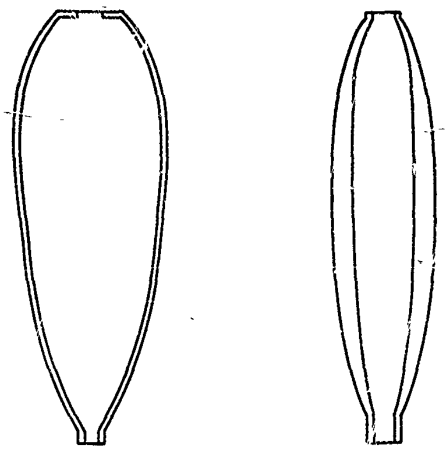
This process is repeated as many times as required to obtain convergence to the minimum weight structure.

The optimum towers for each of the four basic configurations chosen are shown in Figs. 1-6 and 1-7, with a table of results being given in Table 1-1. These results were obtained



(A) One Design Variable *(B) Two Design Variable*

Figure 1-6. Tower With Base Rigidly Fastened to the Earth using a finite element model with approximately forty elements so that the resulting structure has an essentially continuous distribution of material and spacing. The weights shown in Table 1-1, corresponding to no design variables are simply the weights of the optimum towers having uniform members and no variation in spacing. Note that there is a significant reduction in structural weight for the tapered optimum towers over uniform towers. Extensive examples of this kind are presented in Chapters 5, 7, and 9.



(A) One Design Variable *(B) Two Design Variable*

Figure 1-7. Tower With Base Simply Supported and Top Supported With Guy Lines

1-5.2 APPLICATION OF THE STEEPEST DESCENT METHOD IN INTERACTIVE COMPUTER AIDED DESIGN

Very often in design problems, it is not practical to specify a unique cost function to be minimized, hence the formal optimization problem described in par. 1-5.1 does not apply directly. The fact that the vector B in Eq. 1-28 is a direction of steepest descent, however, is extremely valuable information to a designer. The application of this information to a structural design problem, using interactive graphics, is a technique which shows considerable promise in design

Consider, for example, the problem treated in par. 1-5.1. The initial estimate of the optimum tower was taken as a uniform tower. The components of the vector δb can be projected on a cathode ray tube, along with a picture of the structure as shown in Fig. 1-8. The algebraic sign of the components of δb, corresponding to each of the design variables, is an indication of the effect a change in that design variable will have on the cost function

TABLE 1-1

WEIGHTS OF TOWERS

	Cantilevered	Cantilevered	Cantilevered	Guy-line Supported	Guy-line Supported	Guy-line Supported
Number of Design Variables	0	1	2	0	1	2
Best Weight	W = 2440.6 lb	W = 2111.4	W = 1827.9	W = 1563.99	W = 1356.6	W = 1265.71
Height	h = 63.7 in.	h_{max} = 91.4	h_{max} = 80.2	h = 46	h_{max} = 46.5	h_{max} = 36.55
Cross-sectional area of member	A = 7.96 lb	A = 6.97	A_{max} = 10.03	A = 3.84	A = 4.434	A_{max} = 4.95

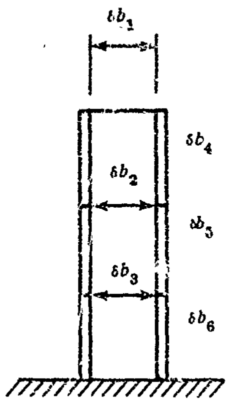


Figure 1-8. Sensitivity to Design Variations

of interest. For example, if δb_1 were positive, this would indicate that an increase in the dimension b_1 will decrease the structural weight. On the other hand if the algebraic sign of δb_1 were negative, then an increase in δb_1 would increase the structural weight. Likewise, the algebraic signs of δb_4 through δb_6 indicate the effect that a change in these element areas will have on structural weight. These data give the designer valuable information, according to which he should change his nominal design to improve the structure, while still satisfying all the essential constraints.

Traditionally, in structural design by graphics, the designer puts areas and dimensions into a structural analysis routine and then requests a stress calculation, the results of which are shown on the screen of a cathode ray tube. This technique has been used by Lockheed-Georgia in the design of the C5A. While this technique has been quite useful in structural design, it is extremely difficult for the designer with only stress information to determine how he should change just one element in the structure to reduce overall structural weight. The difficulty comes in the interplay between structural constraints. If, on the other hand, the designer has trend information that he can use in altering the distribution of material in a structure, he can better use his experience in making design improvements. This capability can be invaluable to large-scale structural designers. It includes the effect of individual design variable changes on overall structural value, while taking into account the effect of that design change on all design constraints.

In real-world structural design problems, the designer must design his structure for more than simply light weight. He must be concerned with structural vibration and buckling characteristics, since these are major sources of structural failure. Often, as in par. 1-5.1, it is possible to determine design perturbations that have a desirable effect on such structural properties as natural frequency and weight simultaneously. Both of these factors can then be displayed on a cathode ray tube as shown in Fig. 1-9. In this case, δb^1 indicates the direction in which the design variable should be changed to reduce structural weight, and δb^2 indicates the direction in which the variable should be changed to increase natural frequency. This information can then be used by experienced design

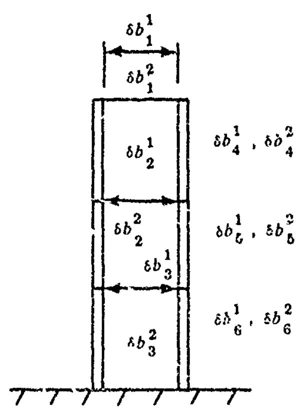


Figure 1-9. Sensitivity to Two Performance Indicators

personnel in making design changes that will have desirable effects on overall aircraft structural properties, for example. This is extremely important in large-scale structural design due to the difficulty in determining the effect of changes in an individual design parameter on several different structural properties. Computation of these data and interactive aspects of the technique are discussed in Chapter 5.

This design technique is feasible from a computational point of view in that very little additional computer time is required to generate sensitivity information from stress and vibration analyses that are required. While most structural optimization work has been done in the batch mode, it is shown in Chapter 5 that utilization of the steepest-descent technique with interactive graphics is a much more practical way to design structures, particularly in cases where several measures of structural performance are important.

Development and display of sensitivity information in design is a form of information transfer to design personnel. This technique depends on the availability of interactive graphics software and hardware, which are currently being developed.

1-5.3 DESIGN OF ARTILLERY RECOIL MECHANISMS

As an application of this same optimization technique to a weapon design problem, certain aspects of the design of a lightweight artillery piece will now be outlined. The requirement was stated for a lightweight artillery piece that can be fired with very short implacement time. For this reason it was determined that the weapon must be capable of being fired while it is resting on its tires. A photograph of the first prototype of this weapon is shown in Fig. 1-10.

The recoil mechanism for this weapon was designed according to traditional recoil mechanism design goals. Namely, the objective in the design was for a constant retarding force which is transmitted by the recoil mechanism to the undercarriage, as shown in Fig. 1-11. A recoil mechanism was designed which delivered approximately this recoil force $R(t)$ as a function of time.

When the weapon was built and fired, a nearly constant recoil occurred, as desired; but, at high angles of fire, the weapon exhibited unacceptable dynamic response. During firing, the tires of the weapon compressed and after firing and the subsequent release of the recoil forces, the weapon rebounded off the ground approximately 6 in. This unacceptable behavior required a redesign cycle for the recoil mechanism with a design goal of minimizing the dynamic response, or hop, of the weapon after firing.

It was determined that the peak recoil force could be allowed to reach 22,000 lb without damaging the support structure. The optimization problem is then to determine the recoil force $R(t)$ as a function time such that

$$R(t) \leq 22{,}000 \qquad (1\text{-}30)$$

and the peak dynamic response, denoted by

$$J = \max_t \,[h(t)] \qquad (1\text{-}31)$$

is as small as possible, where $h(t)$ is the height of the tires off the ground at any time t. Graphically, this problem is to determine a recoil force which lies beneath the 22,000-lb level in Fig 1-12, and which minimizes the peak dynamic response of the weapon. In this

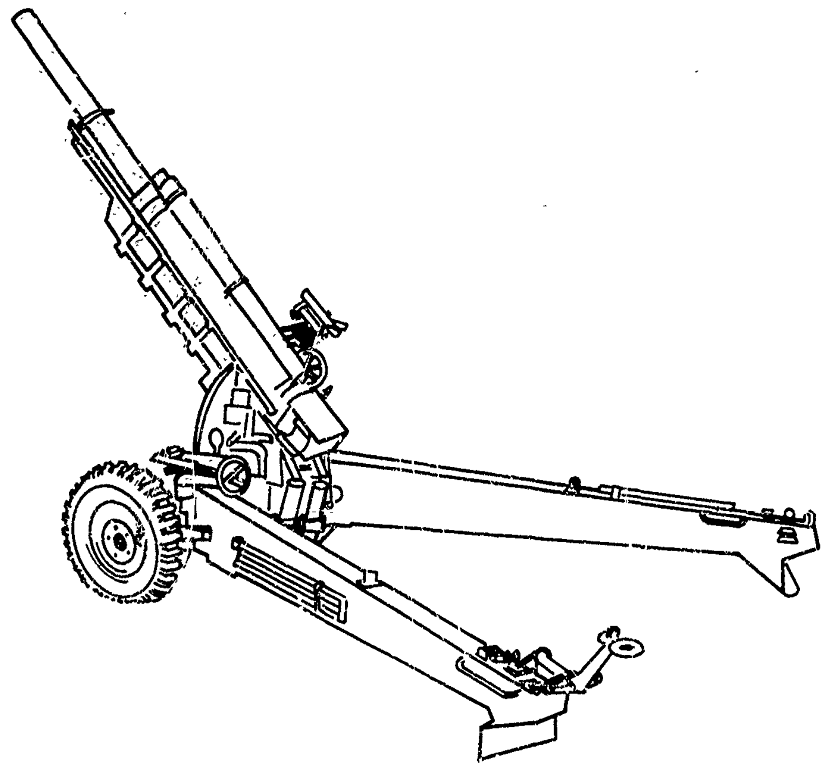


Figure 1-10. Howitzer, Towed, 105 mm, XM164

problem, the dynamic response $h(t)$ is determined by the second order differential equations of motion of the artillery piece.

The same philosophy of small design changes about some nominal estimate, as in the structural design problem of par. 1-5.1, was employed in this case. A sensitivity

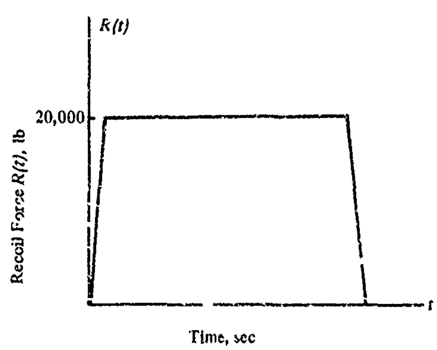


Figure 1-11. Traditional Recoil Design Goal

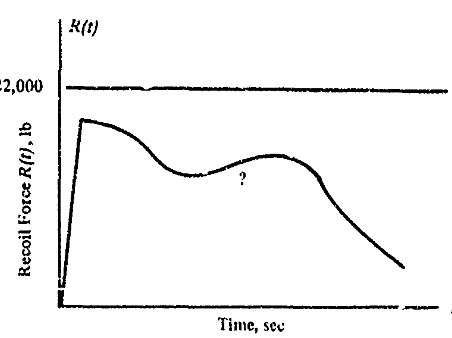


Figure 1-12. Recoil Distribution in Time

function is first determined, which indicates the desirable direction of change in the nominal design variable. For example, taking the previously designed constant retarding force as the nominal base line, a sensitivity function is determined as shown in Fig. 1-13. If a constant multiple of this function is added to the retarding force, a reduction will occur in peak dynamic response and other constraints of the problem will continue to be satisfied. The dotted curve in Fig. 1-13 shows the altered design, which gives better characteristics than the original design estimate.

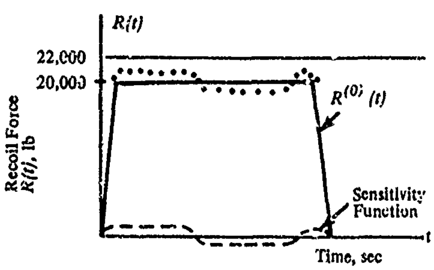


Figure 1-13. Sensitivity to Gun Hop

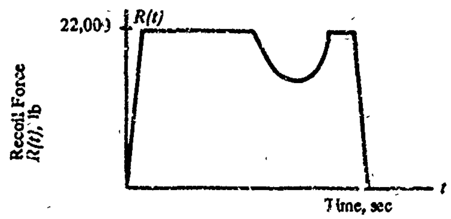


Figure 1-14. Optimum Recoil Curve

This sensitivity information could easily be displayed on the screen of a cathode ray tube and could be used by design personnel in determining desirable changes in the recoil design. Even in this relatively simple problem it was not clear in what way the design should be altered to obtain improved response of the artillery piece. This particular problem was solved in the batch mode by doing many small step iterations of the kind previously described until convergence to an optimum was obtained. The optimum recoil force curve is shown in Fig. 1-14 and resulted in a peak dynamic response of less than 0.5 in. Detailed solution of this problem is presented in Chapter 8, par. 8-5.

REFERENCES

1. *The American College Dictionary*, Random House, New York, 964.
2. C. Y. Sheu and W. Prager, "Recent Developments in Optimal Structural Design", *Appl. Mech. Rev*, Vol. 21, No. 10, October 1968, pp. 985-992.
3. H. H. Goode and R. E. Machol, *System Engineering*, McGraw-Hill, New York, 1957.
4. W. Gosling, *The Design of Engineering Systems*, John Wiley & Sons, New York, 1962.
5. M. K. Starr, *Product Design and Decision Theory*, Prentice-Hall, Englewood Cliffs, New Jersey, 1963.
6. C. Goffman, *Calculus of Several Variables*, Harper and Row, New York, 1965.
7. D. G. Luenberger, *Optimization By Vector Space Methods*, John Wiley & Sons, New York, 1969

CHAPTER 2

FINITE DIMENSIONAL UNCONSTRAINED OPTIMIZATION

2-1 INTRODUCTION

In many engineering design problems certain information which helps to prescribe the object being designed is specified. However, a certain number of parameters called design parameters are left open to the designer's choice. These parameters must uniquely determine the object if the optimal design problem is to be meaningful. In the discussion which follows, the design parameters will be denoted by $x_1 \dots x_n$ or in vector notation simply as $x = (x_1, \dots, x_n)^T$.

In virtually all design problems there are restrictions on the object being designed. These may include the performance required, physical limitations such as [illegible] weight, resource limitations, and o[illegible]al policy. These restrictions or constraints generally will involve the design parameters so that the range of values of design parameters may be restricted. If the vector of design parameters (hereafter called the design parameter) is viewed as an element of real Euclidean space R^n, then the effect of the listed restrictions is to confine the designer's choice of design parameters to a subset D of R^n called the admissible set of design parameters. The nature of this set will be determined by the nature of the requirements placed on the system being designed. This aspect of the optimal design problem will be treated extensively in later chapters.

When one speaks of optimal design, he must be able to choose, out of a collection of objects which satisfy the restrictions of the preceding paragraph, that one which is "best". More specifically, out of all design parameters in the admissible set D, the designer must pick that one, $\bar{x}$, which describes the "best" system. This discussion has still not given the meaning of "best". An effective way of defining "best" is to give a real valued function whose domain of definition is the admissible set D, say $f(x)$. "Best", then, may be taken as the minimum or maximum of $f(x)$ for x in D. If the function $f(x)$ is a cost of the system being designed, then it is to be minimized. If, on the other hand, $f(x)$ is a return or profit, it is to be maximized.

The cost or return function will be defined in each optimal design problem. As a result, very little can be said about its nature in general. It is clear, however, that maximizing a real valued function $r(x)$ is equivalent to minimizing $-r(x)$. Therefore, optimal design problems may always be put into a form which may be interpreted as minimization of a cost function. For convenience this will be done in the following development.

Example 2-1: As a hypothetical optimal design problem let the scalar x be the design parameter and $f(x) = (x - 2)^2$ be the cost function. In Fig. 2-1 the cost function is plotted versus x. It is clear that the minimum cost of zero occurs at $x = 2$.

Example 2-1 is included here as an aid to intuition in more complex problems. Even when x is an n-vector, one can think of plotting the cost function above the x-hyperplane to obtain the cost surface. The optimal design problem is then to find the lowest point on this surface.

Even though real-world optimal design problems invariably have constraints placed on the design parameter, the methods presented in this chapter will ignore constraints. There are two reasons for considering this simplified problem in some detail. First, it may happen that the design parameter $\bar{x}$ that minimizes $f(x)$ lies in the interior of the admissible set D. In this case the constraints play no part in locating $\bar{x}$. Second, even though the point $\bar{x}$ may push some constraint to its limit and lie on the boundary of D, there are iterative methods for finding $\bar{x}$ which require minimization of an auxiliary cost function, subject to no constraints at each iteration. Methods which take constraints into account are presented in Chapters 3 and 4.

Two basically different methods of solving unconstrained minimization problems are presented in this chapter. The first method, called the indirect method, is based on derived properties of the cost function at its minimum; i.e., if one pictures himself as being at the lowest point of the cost surface ($x = 2$ in Fig. 2-1), he may notice that the surface is required to have certain special properties there. He may then use these special properties to locate the lowest point on any such surface. This intuitive idea is made rigorous in par. 2-2.

The second method of solving optimization problems is more direct in nature and is appealing from an engineering point of view. The designer initially chooses a design parameter which is admissible, say $x^{(0)}$. This choice of design parameter will probably not put him at the lowest point on the cost surface. Rather than discarding this nonoptimal point and picking another trial point at random he might attempt to find a second point $x^{(1)}$ which is closer to the lowest point of the cost surface. The designer's view of the cost surface is limited to only a small area due to the local nature of mathematical tests which he may perform. Using only this local information, he chooses a strategy which insures that he makes a move to a new point $x^{(1)}$ which is lower than $x^{(0)}$. The direct methods presented in pars. 2-3 to 2-7 are just a mathematical implementation of these elementary ideas.

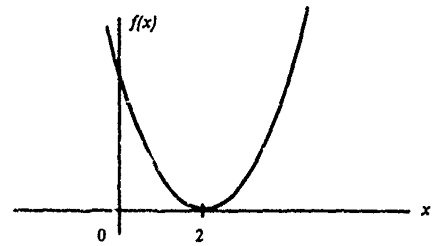


Figure 2-1. $f(x) = (x - 2)^2$

2-2 NECESSARY CONDITIONS FOR EXTREMA

As described in par. 2-1, the approach taken in the indirect method is to assume $f(x)$ has a minimum at $\bar{x}$ and then derive conditions which $f(x)$ must satisfy there. These conditions may then be used to find the minimum point of any real valued function $f(x)$. They are valuable in giving the designer an insight into the minimization portion of an optimal design problem, even when he is using direct computational methods to solve the problem. Before these ideas may be developed, several definitions are required.

Definition 2-1: A real valued function $f(x)$ defined on a subset D of R^n has an absolute minimum at $\bar{x}$ in D if

$$f(\bar{x}) \leqslant f(x) \tag{2-1}$$

for all x in D. The function $g(x)$ has an absolute maximum at $\bar{x}$ if $-g(x)$ has an absolute minimum there. The minimum is called strict if only strict inequalities hold in Eq. 2-1 for $x \neq \bar{x}$.

Note that $f(x)$ can have a strict absolute minimum at only one point in D whereas it could have an absolute minimum at several distinct points in D provided it has the same value at all these points.

Definition 2-2: A function $f(x)$ defined on a subset D of R^n has a relative minimum (maximum) at $\bar{x}$ if there exists an $\epsilon > 0$ so that $f(x)$ has an absolute minimum (maximum) in a subset of D whose points satisfy

$$|x_i - \bar{x}_i| < \epsilon, i = 1, ..., n.$$

Verbally, this definition says that $f(x)$ has a relative minimum at $\bar{x}$ if it has an absolute minimum in a sufficiently small neighborhood of $\bar{x}$. It is clear that if $f(x)$ has an absolute minimum at $\bar{x}$, then it also has a relative minimum there. The converse is not necessarily true.

Example 2-2: Locate all relative and absolute maxima and minima of $f(x)$ on $0 \leqslant x \leqslant 3$, where $f(x)$ is given graphically in Fig. 2-2.

The function $f(x)$ has a strict absolute maximum at $x = 1$, absolute minima (not strict) at $x = 0$ and 2, relative maxima at $x = 1$ and 3, and relative minima at $x = 0$ and 2.

In Definitions 2-1 and 2-2 no continuity or differentiability requirements were placed on

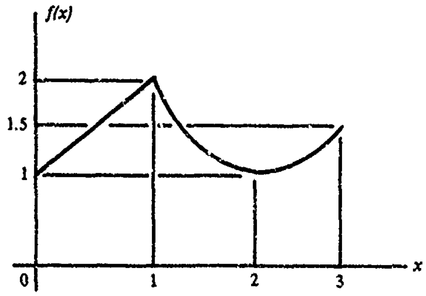


Figure 2-2. A Cost Function

$f(x)$. Without making some assumptions as to the regularity of $f(x)$ it is difficult to verify the required inequalities. Consider the case of a function $f(x)$ of the real variable x which has two continuous derivatives. The Taylor formula is

$$f(\bar{x} + h) = f(\bar{x}) + f'(\bar{x})h + \frac{1}{2}f''(\bar{x} + \theta h)h^2 \tag{2-2}$$

where $0 < \theta < 1$. Since $f''(\bar{x} + \theta h)$ is bounded for h in a closed bounded set, it is clear that if $f'(\bar{x}) \neq 0$ then for small enough h the linear term in h dominates the squared term so that $f(\bar{x} + h)$ may be made both larger and smaller than $f(\bar{x})$ through choice of the appropriate sign of h. Therefore, in order for $f(x)$ to have a relative minimum or maximum at $\bar{x}$ it is necessary that $f'(\bar{x}) = 0$. It follows directly from Eq. 2-2 that if $f'(\bar{x}) = 0$, then $f''(\bar{x}) > 0$ (< 0) is a sufficient condition for $f(x)$ to have a relative minimum (maximum) at $\bar{x}$.

In case x is in R^n, results analogous to those just obtained are given in Theorem 2-1.

Theorem 2-1: Necessary Condition: Let $f(x)$ be defined on a subset D of R^n and have a continuous derivative in a neighborhood of a point $\bar{x}$ which is in the interior of D. If $f(x)$

has a relative minimum at $\bar{x}$ then

$$\nabla f(\bar{x}) = 0 . \tag{2-3}$$

Sufficient Condition: Let $f(x)$ have two continuous derivatives in a neighborhood of $\bar{x}$ and let Eq. 2-3 hold. Then if the matrix

$$\nabla^2 f(\bar{x}) \equiv \left[\frac{\partial^2 f}{\partial x_i \partial x_j} (\bar{x}) \right] \tag{2-4}$$

is positive definite, $f(x)$ has a relative minimum at $\bar{x}$.

For convenience in later discussions, Definition 2-3 is made.

Definition 2-3: A point at which Eq. 2-3 holds is called a stationary point of $f(x)$.

It is imperative that the reader be aware of the hypothesis of Theorem 2-1 which requires $\bar{x}$ to be in the interior of the region D. The theorem does not apply if $\bar{x}$ is on the boundary of D. Example 2-2 illustrates this requirement graphically. Points $x = 0$ and $x = 3$ of Fig. 2-2 yield a relative minimum and a relative maximum, respectively, but neither point is stationary (i.e., neither satisfies Eq. 2-3). The same example also illustrates the need for verification of the differentiability properties of $f(x)$. Even though $x = 1$ yields an absolute maximum of $f(x)$ and is in the interior of D, it is not a stationary point since $f(x)$ does not have a continuous derivative there. This example illustrates the need to faithfully verify all the hypotheses before Theorem 2-1 is employed.

In order to verify the sufficiency condition of Theorem 2-1, one must have a verifiable test for positive definiteness of a matrix. Probably the most useful test is the following (Ref. 2, page 103): A symmetric matrix $A = (a_{ij})_{n \times n}$ is positive definite if and only if the determinate of each of the matrices A_m, formed from the first m rows and first m columns of A, is positive, $m = 1, ..., n$.

Example 2-3: Obtain explicit necessary and sufficient conditions for $f(x_1,x_2)$ to be a minimum and a maximum at $\bar{x}$, where $f(x_1, x_2)$ has two continuous derivatives in D and $\bar{x}$ is an interior point of D.

As necessary conditions for either a minimum or a maximum, Eq. 2-3 demands

$$f_{x_1}(\bar{x}) = f_{x_2}(\bar{x}) = 0 .$$

A sufficient condition for $\bar{x}$ to be a minimum point for $f(x)$ is that in addition to the above equations, the matrices

$$A_1 = f_{x_1x_1}(\bar{x}) \text{ and}$$

$$A_2 = \begin{bmatrix} f_{x_1x_1}(\bar{x}) & f_{x_1x_2}(\bar{x}) \\ f_{x_1x_2}(\bar{x}) & f_{x_2x_2}(\bar{x}) \end{bmatrix}$$

have positive determinates, i.e.,

$$f_{x_1x_1}(\bar{x}) > 0 \text{ and } f_{x_1x_1}(\bar{x}) f_{x_2x_2}(\bar{x}) - [f_{x_1x_2}(\bar{x})]^2 > 0.$$

The function $f(x)$ has a relative maximum at $\bar{x}$ if the function $g(x) = -f(x)$ has a relative minimum there. Therefore, in addition to $-f_{x_1}(\bar{x}) = -f_{x_2}(\bar{x}) = 0$ sufficient conditions for $g(x)$ to have a relative minimum at $\bar{x}$ are

$$g_{x_1x_1}(\bar{x}) > 0 \text{ and } g_{x_1x_1}(\bar{x})\, g_{x_2x_2}(\bar{x}) - [g_{x_1x_2}(\bar{x})]^2 > 0.$$

For a relative maximum of $f(x)$ at $\bar{x}$ then sufficient conditions are

$$f_{x_1 x_1}(\bar{x}) < 0 \text{ and } f_{x_1 x_1}(\bar{x}) f_{x_2 x_2}(\bar{x}) - [f_{x_1 x_2}(\bar{x})]^2 > 0.$$

Thus far in this paragraph only properties of $f(x)$ precisely at the minimum point have been investigated. If the designer viewed the graph of $f(x)$ versus x to be a surface, then Theorem 2-1 tells him what the surface will look like when he finds its lowest point. Theorem 2-1, however, does not tell him that a lowest point exists. In order to solve his optimization problems, the designer would like to have tools which allow him to stand back from the cost surface and learn something about its global properties. Two theorems are now stated which give him a better overall view of the optimization problem.

Theorem 2-2: If $f(x)$ is continuous on a closed and bounded subset D of R^n, then $f(x)$ has an absolute minimum in D.

This theorem does not hold, in general, if any of the hypotheses are deleted. For example, consider the function $f(x) = x$ on $D = (x \mid 0 < x < 1)$. D is not closed and $f(x)$ does not have an absolute minimum in D. If $\bar{D} = \{x \mid 0 \leq x \leq 1\}$ then $\bar{D}$ is closed and $f(x)$ has an absolute minimum at $x = 0$.

Note: The hypotheses of Theorem 2-2 may be weakened by demanding that $f(x)$ be only lower semi-continuous rather than continuous. For proof, see Ref. 1, page 58.

Theorem 2-3 depends on the concept of convexity.

Definition 2-4: A subset D of R^n is called a convex set if whenever x and y are in D, then the straight line segment $x + \theta(y - x)$, $0 \leq \theta \leq 1$, is also in D. A real valued function $f(x)$ defined on a convex set D is called a convex function on D if for any two points y and z in D

$$f[y + \theta(z - y)] \leq f(y) + \theta[f(z) - f(y)], \quad 0 \leq \theta \leq 1.$$

That is, $f(x)$ is convex on D if the straight line segment $f(y) + \theta[f(z) - f(y)]$ is above the graph of $f(x)$ on the line segment $y + \theta(z - y)$ in D, $0 \leq \theta \leq 1$. For a more detailed discussion of convex functions, see Appendix A.

Theorem 2-3 gives the designer valuable information about the global properties of the cost function. It is proved in detail in Chapter 4.

Theorem 2-3: Let $f(x)$ be a convex function defined on a convex set D in R^n. Then a relative minimum of $f(x)$ on D is also an absolute minimum of $f(x)$ on D.

This theorem is of obvious value to the designer. It assures him that if his design problem satisfies the hypotheses of Theorem 2-3 and if he has found a relative minimum then he is through; he has also found the absolute minimum.

Computational methods for finding extrema based on the theorems of this paragraph generally involve the solution of nonlinear algebraic equations. In particular, Eq. 2-3, which is in general nonlinear, can be solved by a numerical method to locate all admissible interior extrema. Methods for solving such equations are given in Ref. 3, Chapter 2. It generally has been found, however, that direct methods for finding extrema are superior to the solution of Eq. 2-3. For this reason no computational methods based on the indirect method will be presented here.

It will be the purpose of the remainder of this chapter to present methods that the designer may use to locate interior relative minima. Relative minima on the boundary of the admissible region will be treated in Chapters 3 and 4.

2-3 ONE-DIMENSIONAL MINIMIZATION

In the direct minimization methods to follow, a multidimensional minimization problem will be reduced to a sequence of one-dimensional minimization problems; i.e., the problem of determining a scalar α so that a given function $g(\alpha)$ will be a minimum.

In the problem of minimizing $f(x)$ for x in R^n, all the methods of solution presented in this chapter are based on successive improvements in certain directions; i.e., at a point $x^{(i)}$ one finds a direction, s, in which $f(x)$ decreases. The object is now to move along the vector $x^{(i)} + \alpha s$, by adjusting α, $\alpha > 0$, until $f(x)$ is as small as possible. The resulting point is then called $x^{(i+1)}$, and the entire process is repeated. It is clear that the intermediate problem of determining α so as to minimize $f(x^{(i)} + \alpha s)$ is one-dimensional. This paragraph will be devoted to presentation of methods for solving the one-dimensional problem.

2-3.1 QUADRATIC INTERPOLATION

If the function $f(x^{(i)} + \alpha s)$ of the scalar variable α – $x^{(i)}$ and the unit vector s are fixed – were quadratic in α, then the value of α which minimizes the function could be found by setting

$$\frac{d}{d\alpha}\left\{ f[x^{(i)} + \alpha s]\right\} = 0.$$

The object here is to treat more general functions, but it is possible to make a quadratic approximation to $f[x^{(i)} + \alpha s]$ which will hold near the minimum point. Then, the minimum point of the approximating function, which may be easily found, is an approximation of the true minimum point.

The quadratic approximation of $f[x^{(i)} + \alpha s]$ is constructed by passing a quadratic curve in α through three computed values of the function. The distance between the three trial points will be $\delta > 0$, where δ is initially chosen to be a small fraction of the expected range of α. It is known, however, that if the starting point of the process is quite far from the minimum point then the minimum point of the approximating function may not be near the true minimum point. To prevent making large, inaccurate steps in this case, a maximum allowable step size Δ is chosen before the process begins. A reasonable choice of Δ is 50% of the expected range of α.

The following algorithm implements the procedure described:

Step 1. Define $\alpha^0 = 0$ and $j = 1$.

Step 2. Compute

$$f_1 = f[x^{(i)} + (\alpha^{j-1} - \delta)s]$$

$$f_0 = f[x^{(i)} + \alpha^{j-1} s]$$

$$f_2 = f[x^{(i)} + (\alpha^{j-1} + \delta)s].$$

Step 3. A quadratic polynomial in $\alpha - \alpha^{j-1} = z$ is fitted through $(-\delta, f_1)$, $(0, f_0)$, (δ, f_2). Its minimum is $z_m = \dfrac{\delta(f_1 - f_2)}{2(f_1 - 2f_0 + f_2)}$, if $f_1 - 2f_0 + f_2 \neq 0$. If this quantity is zero, then the approximation is a straight line with minimum at $z_m = \pm\Delta$, depending on which of f_1 and f_2 is smaller.

Step 4. Define

$$d\alpha = \min(|z_m|,\Delta)\cdot \mathrm{sgn}(z_m)$$

and

$$\alpha^j = \alpha^{j-1} + d\alpha$$

Step 5. If $|d\alpha|$ is less than a specified tolerance, the process is stopped and α^j is taken as the minimum value of α. Otherwise, replace j with $j+1$ and return to Step 2.

2-3.2 FIBONACCI SEARCH (OR GOLDEN SECTION SEARCH)

The Fibonacci search technique is a method based on isolating a relative minimum in an interval and successively decreasing the size of the interval. The process thus gives successively better estimates for the location of the minimum point. For a proof that the method converges very rapidly the reader is referred to Ref. 4, page 236. Here, only the basic ideas behind the method will be given, and an iterative algorithm stated.

Starting at $\alpha = 0$ one might evaluate $f[x^{(i)} + \alpha s]$ at $\alpha = \delta$ and check to see if the functional value is smaller than at $\alpha = 0$. If it is, one might then evaluate the function at $\alpha = 2\delta$ and compare with the value of $\alpha = \delta$. Again if a decrease occurs, one moves on to $\alpha = 3\delta$, etc. The process will terminate when $f[x^{(i)} + (k+1)\delta s] > f[x^{(i)} + k\delta s]$. It is then known that $(k-1)\delta < \alpha < (k+1)\delta$ contains the minimum point and a more accurate result, if required, may be obtained by reducing δ and repeating the process from $\alpha = (k-1)\delta$. If the initial step δ was too small, many steps will have to be made before the minimum point is located.

In Fibonacci search the same basic procedure is followed except that if, after a given step, the functional value has decreased, then the next step size is taken as 1.618 times the previous step size. In this way if the minimum point is a long way from $\alpha = 0$, the Fibonacci technique will isolate it much more rapidly than the previous method with constant step size. Note that there is a penalty, in that the interval which contains the minimum point may have length much greater than 2δ. This is illustrated in Fig. 2-3.

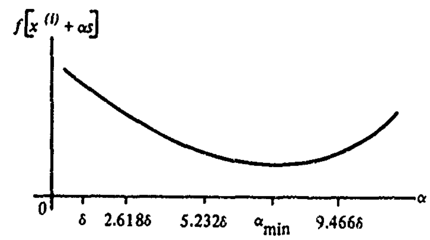


Figure 2-3. Function of Single Variable

Once the minimum point is restricted to some interval, this interval is broken up into three subintervals by inserting points located a distance of 0.382 times the length of the interval from each end. A test is then performed to see which subinterval the minimum point lies in. For a given subinterval the partitioning is shown in Fig. 2-4.

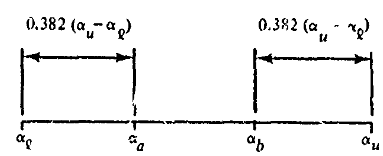


Figure 2-4. Interval Partition

The search process is terminated when the minimum point is isolated in a sufficiently small subinterval.

The Fibonacci search method has the property of being best in a certain sense among all search techniques which isolate α in an interval. A measure of the effectiveness of any such technique is the ratio of the length of the largest interval in which α may lie after n steps to the length of the original interval which contained α. It is shown in Ref. 4, page 253, that if $f[x^{(i)} + \alpha s]$ has a unique relative minimum as a function of α, then Fibonacci search minimizes the number of interval partitions.

The Fibonacci search technique may now be given in the form of a computational algorithm:

Step 1. First an upper bound must be found for α, α_u. It is clear that 0 is a lower bound, α_ℓ. For a chosen small step size δ in α, let j be the smallest integer such that

$$f\left\{x^{(i)} + [\sum_{k=0}^{j} \delta(1.618)^k]s\right\}$$

$$> f\left\{x^{(i)} + [\sum_{k=0}^{j-1} \delta(1.618)^k]s\right\}.$$

Then upper and lower bounds on $\alpha^{(i)}$ are

$$\alpha_u = \sum_{k=0}^{j} \delta(1.618)^k$$

$$\alpha_\ell = \sum_{k=0}^{j-2} \delta(1.618)^k.$$

Step 2. Compute $f[x^{(i)} + \alpha_b s]$, where

$$\alpha_a = \alpha_\ell + 0.382(\alpha_u - \alpha_\ell)$$

$$\alpha_b = \alpha_\ell + 0.618(\alpha_u - \alpha_\ell).$$

Note that $\alpha_a = \sum_{k=0}^{j-1} \delta(1.618)^k$ so $f[x^{(i)} + \alpha_a s]$ is known.

Step 3. Compare $f[x^{(i)} + \alpha_a s]$ and $f[x^{(i)} + \alpha_b s]$ and go to Step 4, 5, or 6.

Step 4. If $f[x^{(i)} + \alpha_a s] < f[x^{(i)} + \alpha_b s]$, then $\alpha_\ell < \alpha^{(i)} < \alpha_b$. By the choice of α_a and α_b, the new points $\alpha'_\ell = \alpha_\ell$ and $\alpha'_u = \alpha_b$ have $\alpha'_b = \alpha_a$. Compute now $f[x^{(i)} + \alpha'_a s]$ where $\alpha'_a = \alpha'_\ell + 0.382(\alpha'_u - \alpha'_\ell)$. Go to Step 7.

Step 5. If $f[x^{(i)} + \alpha_a s] > f[x^{(i)} + \alpha_b s]$, then $\alpha_a < \alpha^{(i)} < \alpha_u$. Similar to the procedure in Step 4, put $\alpha'_\ell = \alpha_a$ and $\alpha'_u = \alpha_u$ so that $\alpha'_a = \alpha_b$. Compute $f[x^{(i)} + \alpha'_b s]$ where $\alpha'_b = \alpha'_\ell + 0.618(\alpha'_u - \alpha'_\ell)$. Go to Step 7.

Step 6. If $f[x^{(i)} + \alpha_a s] = f[x^{(i)} + \alpha_b s]$, put $\alpha'_\ell = \alpha_a$ and $\alpha'_u = \alpha_b$. Return to Step 2.

Step 7. If $\alpha'_u - \alpha'_\ell$ is suitably small, put $\alpha^{(i)} = \frac{1}{2}(\alpha'_u + \alpha'_\ell)$ and stop. Otherwise, delete the primes on α'_ℓ, α'_a, α'_b, and α'_u and return to Step 3

2-4 THE METHOD OF STEEPEST DESCENT (OR GRADIENT)

The simplest and probably the best known of the direct methods of minimization is the Method of Steepest Descent (or Gradient). This method is based on the fact that if the cost surface is smooth, then its tangent plane is a good approximation to the surface near the point of tangency. The philosophy of the

Method of Steepest Descent is apparent in its title. One wishes to change $x^{(i)}$ by an increment dx in such a way that $f(x)$, $x = x^{(i)} + dx$, is reduced as much as possible for a given length of increment. The direction of the increment dx is called the direction of steepest descent.

The direction of steepest descent is given by Theorem 2-4.

Theorem 2-4: Let $f(x)$ be differentiable in R^n. The direction of steepest descent at a point $\tilde{x}$ is

$$dx = -\alpha \nabla f^T(\tilde{x}) \tag{2-4}$$

where $\alpha > 0$ is a scalar factor.

The proof of Theorem 2-4 illustrates clearly that the direction of steepest ascent is

$$dx = \alpha \nabla f^T(x) \tag{2-5}$$

for $\alpha > 0$. The reader should note carefully that Eqs. 2-4 and 2-5 give only the direction in the design parameter space R^n which yields the maximum rate of change of $f(x)$. Since the factor α is not determined explicitly, the size of step is not specified.

In order to start the steepest descent iterative technique, the designer makes the best estimate of the design parameter available, $x^{(0)}$. The gradient $\nabla f[x^{(0)}]$ is then computed at $x^{(0)}$ and a new point $x^{(1)}$ is determined by

$$x^{(1)} = x^{(0)} - \alpha^{(0)} \nabla f^T[x^{(0)}]$$

where $\alpha^{(0)} > 0$ is chosen using methods of par. 2-3 so that $f[x^{(0)} - \alpha \nabla f^T(x^{(0)})]$ is a minimum as a function of α. If $\nabla f[x^{(0)}] \neq 0$ then $f[x^{(1)}] < f[x^{(0)}]$, so $x^{(1)}$ is taken as a better estimate of the minimum point and the process is continued until $\nabla f[x^{(i)}] = 0$ or dx is sufficiently small. This method may be given in compact form as the steepest descent algorithm:

Step 1. Make the best engineering estimate $x^{(0)}$ of the minimum point.

Step 2. Compute $\nabla f[x^{(i)}]$ and define a normalized gradient $s = \dfrac{1}{\| \nabla f[x^{(i)}] \|} \nabla f^T[x^{(i)}]$. Find $\alpha = \alpha^{(i)}$ which minimizes $f[x^{(i)} + \alpha s]$ (where i is the number of iterations completed). If $\nabla f[x^{(i)}] = 0$, terminate the process and $x^{(i)}$ is a relative minimum point.

Step 3. Put $x^{(i+1)} = x^{(i)} - \alpha^{(i)} s$. If $|\alpha^{(i)}|$ and $\| \nabla f[x^{(i+1)}] \|$ are less than predetermined limits, terminate the process and let $x^{(i+1)}$ be the approximation to the minimum point. Otherwise return to Step 2.

It is interesting to note that successive directions of steepest descent are orthogonal to one-another in this algorithm — i.e., $\nabla f[x^{(i+1)}]\ \nabla f^T[x^{(i)} = 0]$. To see this, recall that $\alpha^{(i)}$ is chosen so that $f[x^{(i)} - \alpha s]$ is a minimum in α. The necessary condition of Theorem 2-1 then requires

$$0 = \frac{\partial f}{\partial \alpha} = -\frac{1}{\| \nabla f[x^{(i)}] \|} \sum_{j=1}^{n} \frac{\partial f}{\partial x_j}[x^{(i+1)}]$$
$$\frac{\partial f}{\partial x_j}[x^{(i)}] = -\frac{1}{\| \nabla f^T[x^{(i)}] \|}$$
$$\nabla f[x^{(i+1)}]\ \nabla f^T[x^{(i)}]$$

which was to be shown.

In the case where $x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$, Fig. 2-5 is a view of the design variable space. The closed

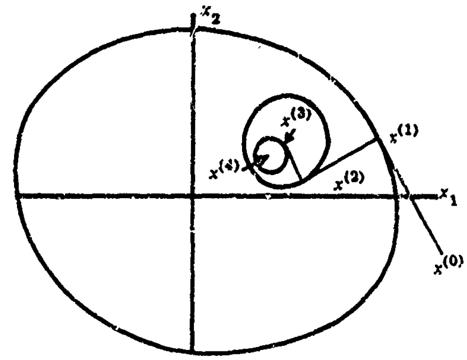


Figure 2-5. Descent Steps

curves in this figure are lines of constant $f(x)$.

A relatively general convergence theorem pertaining to this algorithm will now be stated. The proof of this theorem may be found in Ref. 5, page 80.

Theorem 2-5: Let $f(x)$ be a continuous function defined on R^n and let $x^{(0)}$ be any point such that the closed set

$$S = \left\{ x \mid f(x) < f[x^{(0)}] \right\}$$

is bounded, and $f(x)$ is twice continuously differentiable on S. Let the matrix of second derivatives of $f(x)$,

$$H(x) = \left[\frac{\partial^2 f(x)}{\partial x_i \partial x_j}\right]$$

satisfy the condition

$$| y^T H y | < M y^T y$$

for some M, every y in R^n, and every x in S. Then for the sequence $[x^{(i)}]$ generated by the steepest descent algorithm:

(1) A subsequence $x^{(i_m)}$ converges to a point $\bar{x}$ in S for which $\nabla f(\bar{x}) = 0$.

(2) $f[x^{(i_m)}]$ decreases monotonically to $f(\bar{x})$.

(3) If $\bar{x}$ is the only point in S for which $\nabla f(\bar{x}) = 0$, then $x^{(i)}$ converges to $\bar{x}$.

Several things which Theorem 2-5 does *not* say are worthy of note. First, the theorem does not guarantee that the sequence of points $x^{(i)}$ generated by the Method of Steepest Descent will converge. Further, unless the assumption of (3) holds, the sequence need not converge to an absolute minimum; it may converge to a relative minimum.

The choice of the initial estimate $x^{(0)}$ can have a great deal to do with the limit point of the algorithm if it does converge. If it is not known beforehand that a unique relative minimum exists, it is general practice to start the iterative process at several initial estimates. If the sequence $x^{(i)}$ converges to the same point $\bar{x}$ each time, then one is led to believe that he has indeed found an absolute minimum. Logic such as this can cause sleepless nights, however, particularly if a decision involving considerable resources and perhaps even one's job depends on the outcome. For this reason, the importance of at least making a serious attempt to apply theorems such as those of par. 2-2 cannot be overemphasized. Theorem 2-3, for example, if properly applied, may prevent many anxious moments.

In spite of the simplicity of the Method of Steepest Descent, it has several severe restrictions which are discussed in Ref. 5, page 159. These are:

1. Even though convergence may be guaranteed by Theorem 2-5, an infinite number of iterations may be required for the minimization of even a positive definite quadratic form.

2. Each iteration is calculated independently of the others so that no information is

stored which might be used to accelerate convergence

3. The rate of convergence depends strongly on properties of the cost function. If the ratio of the largest and smallest eigenvalues of the matrix of second derivatives is large, the steepest descent algorithm generates short zig-zagging moves. Convergence is, therefore, very slow.

For an extensive treatment of modifications of the steepest descent method, which prevents certain of these difficulties, see Ref. 4, Chapter 7. Several methods, presented in the next three paragraphs, do not suffer so severely from the problems just described.

2-5 A GENERALIZED NEWTON METHOD

In the Steepest Descent Method of par. 2-4, only first-order derivatives that determine the tangent plane of the cost surface are used to represent the behavior of this surface. One would expect that if second derivatives of the cost function were available, then a quadratic function could be constructed as an approximation to the surface. The quadratic approximation should allow for much better approximation of the minimum point of the cost function.

The idea of this method is to first use a second-degree Taylor formula as an approximation to $f(x)$. If $f(x)$ is convex, or just convex near a minimum point then the minimum point of the quadratic should be near the minimum point of $f(x)$. The minimum point of the quadratic approximation is then determined analytically and is taken as a good approximation of the minimum point of $f(x)$.

In order to utilize Taylor's formula including second degree terms, the following matrix is required

$$H(x) \equiv \nabla^2 f(x) \equiv \left[\frac{\partial^2 f(x)}{\partial x_i \partial x_j}\right]_{n \times n}$$

Note that it is implicitly assumed here that $f(x)$ has two derivatives. By Taylor's formula,

$$f[x^{(0)} + \Delta x] \approx f[x^{(0)}] + \nabla f[x^{(0)}]\Delta x + \frac{1}{2}\Delta x^T H[x^{(0)}]\Delta x \quad (2\text{-}6)$$

where Δx is a change in $x^{(0)}$. In case $f(x)$ is locally convex – convex in a neighborhood of $x^{(0)}$ – Theorem A-3 shows that $H[x^{(0)}]$ is positive semi-definite. If, in addition, $H[x^{(0)}]$ is positive definite, then it has an inverse. Further, $f[x^{(0)} + \Delta x]$ in Eq. 2-6 is convex in Δx so Theorem 2-3 insures the existence of a unique minimum point of the quadratic function in Eq. 2-6. By Theorem 2-1, this unique minimum point is determined by

$$\nabla f^T[x^{(0)}] + H[x^{(0)}]\Delta x = 0$$

or

$$\Delta x = -H^{-1}[x^{(0)}]\ \nabla f^T[x^{(0)}], \quad (2\text{-}7)$$

and the new estimate of the minimum point is $x^{(1)} = x^{(0)} + \Delta x$.

Since Eq. 2-6 is just an approximation, $x^{(1)}$ will probably not be the precise minimum point of $f(x)$. Realizing that evaluation of $H(x)$ requires computation of $n(n + 1)/2$ second derivatives of $f(x)$, one might be tempted to improve the estimate for the minimum point before recalculating all these derivatives.

An easy way of improving the estimate of the minimum point is to change the length of the step Δx without altering its direction. The scalar $\alpha \simeq 1$ will be determined by methods of par. 2-3 so as to minimize $f\,[x^{(0)} + \alpha\Delta x]$.

This procedure may now be put down in the form of a computational algorithm called Generalized Newton Method:

Step 1. Make an engineering estimate $x^{(0)}$ of the minimum point of $f(x)$.

Step 2. Compute

$$x^{(i+1)} = x^{(i)} - \alpha^{(i)} H^{-1}[x^{(i)}]\ \nabla f^T[x^{(i)}]\ ,$$

where $\alpha = \alpha^{(i)}$ is chosen which minimizes

$$f\left\{x^{(i)} - \alpha H^{-1}[x^{(i)}]\ \nabla f^T[x^{(i)}]\right\}$$

as a function of α. Here, the index i is the number of iterations completed.

Step 3. If $\|\nabla f\,[x^{(i)}]\|$ and $\|x^{(i+1)} - x^{(i)}\|$ are sufficiently small, terminate the process and take $x^{(i+1)}$ as the minimum point of $f(x)$. Otherwise, return to Step 2.

The Generalized Newton Method presented in this paragraph has been called the best for minimizing convex cost functions when second derivatives are available (see Ref. 5, page 162). Even in the case in which the cost function is nonconvex, if the starting point $x^{(0)}$ is near enough to a relative minimum point so that the cost function is convex at $x^{(0)}$, then good convergence may still be expected.

In spite of the advantages of this method, it still has several shortcomings.

1. Even if $f(x)$ is convex an inverse of $H(x)$ may not exist unless $H(x)$ is strictly positive definite.

2. In nonconvex problems an iteration does not necessarily decrease $f[x^{(i)}]$ when the current iterate $x^{(i)}$ is not near the minimum point.

3. For many engineering problems, $H(x)$ will be extremely messy if not impossible to compute efficiently.

Even in nonconvex minimization problems the Generalized Newton Method may be used in conjunction with a Steepest Descent Method to form an extremely effective tool. The Steepest Descent Method has the property of making good progress even though only a poor estimate of the minimum point is available. As a relative minimum is approached, however, the rate of convergence of the Steepest Descent Method decreases. At this point, however, the cost function should be convex since a minimum point is nearby. Therefore, the Generalized Newton Method may be employed for rapid convergence to the relative minimum point.

2-6 METHODS OF CONJUGATE DIRECTIONS

In par. 2-4 it is pointed out that the Method of Steepest Descent had rather poor convergence properties in many problems because it uses only first-order approximations (involving only first-order derivatives). Further, the Steepest Descent Method is not a learning process in that it does not store information from past iterations. The first deficiency is corrected in par. 2-5 where a Generalized Newton Method employing second derivatives is presented. This method, while having outstanding convergence properties, requires the computation of $n(n + 1)/2$

second-order derivatives at each iteration (x is in R^n). In most engineering design problems this is an extremely tedious, if not impossible, task. Further, the Generalized Newton Method is not a learning process.

The methods presented in this paragraph require the computation of only first derivatives. However, by making use of information obtained from previous derivatives, convergence is speeded as the minimum is approached. In fact, as one of the methods progresses, it develops an approximation to the matrix of second derivatives. In this respect the methods here have the desirable features of both the Method of Steepest Descent and the Generalized Newton Method.

All Methods of Conjugate Directions are based on the philosophy "If a method works well in minimizing all positive definite quadratic forms, then it ought to work pretty well on any smooth cost function." To be more specific, Conjugate Gradient Methods are guaranteed to minimize any positive definite quadratic form in n iterations (the design parameter is in R^n). Although this ideal behavior will not carry over to general cost functions, since a convex cost function often looks very much like a positive definite quadratic form, similar behavior could be expected. Experience has shown that this is the case.

In order to be more precise, one makes Definition 2-5.

Definition 2-5: Let A be a symmetric positive definite $n \times n$ matrix and S^i, $i = 0, 1, ..., n-1$, be nonzero vectors in R^n. The S^i are called conjugate with respect to A if

$$S^{i^T} A S^j = 0, \; i \neq j \qquad (2\text{-}8)$$

Since A is positive definite, the conjugate vectors S^i are linearly independent. To see that this is true, form the linear combination

$$\sum_{i=0}^{n-1} a_i S^i = 0,$$

where the a_i are scalars. Multiplying this sum on the left by $S^{j^T} A$ yields

$$\sum_{i=0}^{n-1} a_i S^{j^T} A S^i = a_j S^{j^T} A S^j = 0$$

and since $S^{j^T} A S^j \neq 0$, $a_j = 0$. Since j was arbitrary, $a_j = 0$, $j = 0, 1, ..., n-1$, and this is just the definition of linear independence.

Consider now the problem of minimizing the convex function

$$f(x) = B^T x + \frac{1}{2} x^T A x \qquad (2\text{-}9)$$

where x is in R^n, B is an $n \times 1$ matrix and A is a symmetric positive definite, $n \times n$ matrix. The central idea of all methods based on conjugate directions is contained in Theorem 2-6.

Theorem 2-6: Let $S^0, ..., S^{n-1}$ be nonzero vectors in R^n which are conjugate with respect to the positive definite matrix A. Choose scalars $\lambda = \lambda^{(i)}$, $i = 0, ..., n-1$, successively which minimize

$$f[x^{(i)} + \lambda S^i] \qquad (2\text{-}10)$$

where $f(x)$ is given in Eq. 2-9,

$$x^{(i)} = x^{(0)} + \sum_{k=0}^{i-1} \lambda^{(k)} S^k \qquad (2\text{-}11)$$

and $x^{(0)}$ is any point in R^n. Then $x^{(n)}$ is the absolute minimum point of $f(x)$ over R^n.

The two methods that follow are simply

based on different ways of generating conjugate directions. There are an unlimited number of ways to generate conjugate directions. Several ways are discussed in Ref. 6.

2-6.1 THE CONJUGATE GRADIENT METHOD

Given any set of n linearly independent vectors and a positive definite $n \times m$ matrix A a set of conjugate directions with respect to A can be generated by a Gram-Schmidt orthogonalization technique. Let v^0, ..., v^{n-1} be linearly independent vectors and define $S^0 = v^0$. Now put

$$S^1 = v^1 + \alpha_{10} S^0.$$

For A-conjugacy, it is required that

$$S^{0T} A S^1 = 0 = S^{0T} A (v^1 + \alpha_{10} S^0)$$

and

$$\alpha_{10} = -\frac{v^{1T} A S^0}{S^{0T} A S^0}.$$

Assuming S^1, ..., S^k are A-conjugate, put

$$S^{k+1} = v^{k+1} + \alpha_{k+1,0} S^0 + \ldots + \alpha_{k+1,k} S^k.$$

For A-conjugacy it is required that

$$S^{k+1\,T} A\, S^r = 0 = v^{k+1\,T} A\, S^r + \alpha_{k+1,r} S^{rT} A\, S^r$$

where the second equality holds by S-conjugacy, so

$$\alpha_{k+1,r} = -\frac{v^{k+1\,T} A\, S^r}{S^{rT} A\, S^r}, r = 1, \ldots, k.$$

By induction, the resulting directions are A-conjugate and

$$S^{k+1} = v^{k+1} - \sum_{r=0}^{k} \frac{v^{k+1\,T} A\, S^r}{S^{rT} A\, S^r}.$$

Many sets of vectors v_i could be chosen to generate conjugate directions. A natural choice, however, is the set of gradient vectors of $f(x)$, $g^i = \nabla f^T(x^{(i)})$, where $x^{(i)}$ are defined in Theorem 2-6. Define

$$S^0 = -g^0$$

$$S^{k+1} = -g^{k+1} + \sum_{i=0}^{k} \frac{g^{k+1\,T} A\, S^i}{S^{iT} A\, S^i} S^i. \quad (2\text{-}12)$$

Alternatively,

$$g^{k+1} = -S^{k+1} + \sum_{i=0}^{k} \frac{g^{k+1\,T} A\, S^i}{S^{iT} A\, S^i} S^i. \quad (2\text{-}13)$$

Since $f(x) = \frac{1}{2} x^T A x + B^T x$,

$$g_k = \nabla f^T [x^{(k)}] = A x^{(k)} + B,$$

or from the proof of Theorem 2-6,

$$g_k = g^{i+1} + A \left[\sum_{\ell=i+1}^{k-1} \lambda^{(\ell)} S^\ell \right]. \quad (2\text{-}14)$$

Now,

$$g^{kT} S^i = g^{i+1\,T} S^i + S^{iT} A \left[\sum_{\ell=i+1}^{k-1} \lambda^{(\ell)} S^\ell \right]$$

$$= 0, i < k \quad (2\text{-}15)$$

due to A-conjugacy of the S^i and

$$\nabla f [x^{(k+1)}] S^k = 0, k = 0, \ldots, n-1. \quad (2\text{-}16)$$

From Eqs. 2-13 and 2-14

$$g^{kT} g^i = g^{kT} \left[-S^i + \sum_{j=0}^{i-1} \frac{g^{iT} A\, S^j}{S^j A\, S^j} S^j \right]$$

$$= 0, i \quad k \quad (2\text{-}17)$$

Thus, the g^i, $i = 0, 1, ..., n-1$ are linearly independent and the S^i, $i = 0,1, ..., n-1$ are A-conjugate.

The Conjugate Direction Method of Theorem 2-6 may now be applied using the conjugate gradients S^i. The result is called the Conjugate Gradient Method. In order to apply this method to nonquadratic problems, it is first necessary to eliminate explicit dependence of the algorithm on the form of $f(x)$.

By definition,

$$g^{j+1} = Ax^{j+1} + B = A\,[x^j + \lambda^{(j)} S^j] + B$$

or

$$g^{j+1} = g^j + \lambda^{(j)} AS^j. \tag{2-18}$$

By Eq. 2-16

$$g^{j+1^T} S^j = 0 = g^{j^T} S^j + \lambda^{(j)} S^{j^T} A\, S^j.$$

Thus,

$$\lambda^{(j)} = -\frac{g^{j^T} S^j}{S^{j^T} A\, S^j}.$$

Substituting for S^j from Eq. 2-12 and using Eq. 2-15, this is

$$\lambda^{(j)} = \frac{g^{j^T} S^j}{S^{j^T} A\, S^j}. \tag{2-19}$$

From Eqs. 2-18 and 2-19

$$A\, S^i = \frac{S^{i^T} A\, S^i}{g^{i^T} g^i}\,(g^{i+1} - g^i).$$

Now,

$$\frac{g^{k+1^T} A\, S^i}{S^{i^T} A\, S^i} = \frac{g^{k+1^T}(g^{i+1} - g^i)}{g^{i^T} g^i}$$

By Eq. 2-17, for $i < k$, the right side of the above equation is zero. For $i = k$,

$$\frac{g^{k+1^T} A\, S^k}{S^{k^T} A\, S^k} = \frac{g^{k+1^T} g^{k+1}}{g^{k^T} g^k}.$$

Substituting this result into Eq. 2-12 yields

$$S^{k+1} = -g^{k+1} + \left(\frac{g^{k+1^T} g^{k+1}}{g^{k^T} g^k}\right) S^k. \tag{2-20}$$

Eq. 2-20 now gives an algorithm for determining the conjugate directions, even without knowledge of the matrix A.

For a general function $f(x)$,

$$g^i = \nabla f^T [x^{(i)}]$$

and the following algorithm for finding the unconstrained minimum of $f(x)$ is called the *Conjugate Gradient Method:*

Step 1. Make an engineering estimate $x^{(0)}$ of the minimum point and compute

$$S^0 = -\nabla f^T [x^{(0)}]$$

Step 2. For $i = 0,1, ..., $ find $\alpha = \alpha^{(i)}$ which minimizes $f\,[x^{(i)} + \alpha S^i]$.

Step 3. Compute

$$x^{(i+1)} = x^{(i)} + \alpha^{(i)} S^i$$

$$S^{i+1} = -\nabla f^T [x^{(i+1)}] + \beta^i S^i$$

where

$$\beta^i = \frac{\nabla f[x^{(i+1)}]\, \nabla f^T [x^{(i+1)}]}{\nabla f[x^{(i)}]\, \nabla f^T [x^{(i)}]}.$$

Step 4. Terminate the process if $\|\nabla f[x^{(i+1)}]\|$ and $\|x^{(i+1)} - x^{(i)}\|$ are sufficiently small. Otherwise, return to Step 2.

When this algorithm is applied to problems in which $f(\cdot)$ is not of the form of Eq. 2-9, convergence will not occur in n steps. Fletcher and Reeves recommend that after n steps the algorithm should be "restarted", i.e., $x^{(n+1)}$ should be treated as $x^{(0)}$ in the algorithm. In a sense, the first few iterations of the algorithm build up information about the curvature of the cost surface. After n iterations, this information is discarded and a new estimate of curvature is built up during the next n iterations. This method then does not accumulate information about curvature of the cost surface over the entire iterative process.

2-6.2 THE METHOD OF FLETCHER AND POWELL

A second method based on a different set of conjugate directions was suggested by Davidon (Ref. 8) and modified by Fletcher and Powell (Ref. 9). This method is reported to be one of the most powerful known methods for general functions $f(x)$, (Ref. 10). A major reason for the success of this method is its capability to accumulate information about the curvature of the cost surface during the entire iterative process, even though only first order derivatives of the cost function need to be computed.

The directions $S^{(i)}$, generated by the algorithm that follows, are conjugate if $f(x)$ is of the form of Eq. 2-9. This proof is given in Refs. 7 and 9. In Ref. 6 it is shown that the method of Fletcher and Powell fits naturally into a large class of conjugate direction methods. The derivation is tedious and lends little insight into use of the method. For a direct proof of convergence, etc., the reader is referred to Ref. 7.

The computational algorithm is:

Step 1. Make an engineering estimate $x^{(0)}$ of the minimum point and choose a symmetric positive definite matrix $H^{(0)}$.

Step 2. For $i = 0, \ldots,$ compute

$$S^{(i)} = -H^{(i)}\nabla f^T[x^{(i)}].$$

Step 3. Compute $\alpha = \alpha^{(i)}$ which minimizes $f[x^{(i)} + \alpha S^{(i)}]$.

Step 4. Compute

$$\sigma^{(i)} = \alpha^{(i)} S^{(i)}$$

$$x^{(i+1)} = x^{(i)} + \sigma^{(i)}$$

$$H^{(i+1)} = H^{(i)} + A^{(i)} + B^{(i)}$$

where

$$y^{(i)} = \nabla f^T[x^{(i+1)}] - \nabla f^T[x^{(i)}]$$

$$A^{(i)} = \frac{\sigma^{(i)}\sigma^{(i)T}}{\sigma^{(i)T}y^{(i)}}$$

$$B^{(i)} = -\frac{H^{(i)}y^{(i)}y^{(i)T}H^{(i)}}{y^{(i)T}H^{(i)}y^{(i)}}.$$

Step 5. If $\|\nabla f[x^{(i+1)}]\|$ and $\|x^{(i+1)} - x^{(i)}\|$ are sufficiently small, terminate the process. Otherwise return to Step 2.

Fletcher and Powell (Ref. 9) prove that this algorithm has the following properties:

1. The matrix $H^{(i)}$ is positive definite for all i. This implies the method will always converge to a stationary point since

$$\frac{d}{d\alpha} f[x^{(i)} + \alpha S^{(i)}]\Big|_{\alpha=0}$$

$$= -\nabla f[x^{(i)}] H^{(i)} \nabla f^T [x^{(i)}] < 0$$

provided $\nabla f[x^{(i)}] \neq 0$. This means that $f[x^{(i)}]$ may be decreased by choosing $\alpha > 0$ if $f[x^{(i)}] \neq 0$.

2. When this method is applied to the positive definite quadratic from Eq. 2-9, $H^{(i)}$ converges to A^{-1}.

This method might be called a learning process in that only first derivatives are ever computed, but as the algorithm progresses an approximation of the matrix of second derivatives is generated. Many experienced researchers in the area of optimization methods laud this method as one of the best available.

2-6.3 A CONJUGATE DIRECTION METHOD WITHOUT DERIVATIVES

Occasionally in applications, one is faced with a problem in which computation of derivatives of the cost function is impossible or at least prohibitive from a computational point of view. There are many techniques for solving this sort of problem given in Ref. 4. An efficient technique, not presented in common texts, was developed by Powell (Ref. 11) using conjugate directions.

A computational algorithm is presented here without proof. For a proof that the algorithm generates conjugate directions the reader is referred to Ref. 11. The computational algorithm is:

Step 1. Make an engineering estimate of the minimum point $x^{(0)}$ of $f(x)$. Choose vectors s^j, $j = 1, ..., n$, in the coordinate directions of R^n.

Step 2. Find $\alpha = \alpha^k$, $k = 1, ..., n$, which minimize $f[x^{(k-1)} + \alpha s^k]$

where

$$y^0 = x^{(i)}$$

$$y^k = y^{k-1} + \alpha^k s^k,\ k = 1, ..., n,$$

and i is the number of iterations which have been completed. Note that in the one dimensional minimization for α^k, it is possible for $\alpha^k < 0$.

Step 3. Find the integer m, $1 \leq m \leq n$ for which

$$f(y^{m-1}) - f(y^m)$$

is the largest and define

$$\Delta = f(y^{m-1}) - f(y^m).$$

Step 4. Define $f_1 = f(y^0)$, $f_2 = f(y^n)$, and $f_3 = f(2y^n - y^0)$.

Step 5. If $f_3 \geq f_1$ or

$$(f_1 - 2f_2 + f_3) \times (f_1 - f_2 - \Delta)^2$$

$$\geq \frac{\Delta}{2}(f_1 - f_3)^2,$$

put

$$x^{(i+1)} = y^n.$$

Terminate the process if $\|x^{(i+1)} - x^{(i)}\|$ is sufficiently small. Otherwise return to Step 2 with the same set of s^j, $j = 1, \ldots, n$.

Step 6. If neither of the inequalities of Step 5 hold, define $s = y^n - y^0$) and find $\alpha = \bar{\alpha}$ which minimizes

$$f(y^n + \alpha s).$$

Put:

$$x^{(i+1)} = y^n + \bar{\alpha} s.$$

Terminate the process if $\|x^{(i+1)} - x^{(i)}\|$ is sufficiently small. Otherwise return to Step 2 with the new set of vectors $x^1, \ldots, x^{m-1}, s^{m+1}, \ldots, s^n, s$.

For a discussion of use of the algorithm, see Ref. 11.

2-7 COMPARISON OF THE VARIOUS METHODS

During the development of the methods presented in this chapter, theoretical advantages and disadvantages have been pointed out. As a concrete test of these methods, three functions will be minimized. Two of the functions to be treated are terribly behaved and pose a meaningful test to any general minimization technique. These functions resemble a very deep valley at whose bottom the curvature in two orthogonal directions is radically different. The third function is quadratic and poses no serious obstacle to any reasonable method. More specifically, these functions are

$$f_1(x_1, x_2) = 100(x_2 - x_1^2)^2 + (1 - x_1)^2 \tag{2-21}$$

$$f_2(x_1, x_2, x_3, x_4) = (x_1 + 10x_2)^2 + 5(x_3 - x_4)^2 + (x_2 - 2x_3)^4 + 10(x_1 - x_4)^4 \tag{2-22}$$

and

$$f_3(x_1, x_2, x_3) = x_1^2 + 2x_2^2 + 2x_3^2 + 2x_1x_2 + 2x_2x_3 \tag{2-23}$$

The reader should verify that each of these functions has a strict absolute minimum point. These points are (1,1), (0,0,0,0), and (0,0,0), respectively. Each iterative method will be started at points (– 1,1), (1,1,1,1), and (1,1,1) for Eqs. 2-21, 2-22, and 2-23, respectively. These functions will all be minimized by each of the methods of pars. 2-4 through 2-6. The stopping criterion will be that each component of the independent variable must be within 10^{-2} of the known minimum point.

Results will be presented in tabular form so that a comparison of the behavior of each of the methods may be made. For the sake of uniformity, each table will include the iteration number i, the iterate $x^{(i)} = [x_1^{(i)}, \ldots, x_n^{(i)}]^T$, and the value of the cost function.

2-7.1 METHOD OF STEEPEST DESCENT

2-7.1.1 COST FUNCTION: $f_1(x) = 100(x_2 - x_1^2)^2 + (1 - x_1)^2$.

Exact solution: $(1,1)$, $f_1(1,1) = 0$

TABLE 2-1

STEEPEST DESCENT METHOD – ITERATIVE DATA FOR COST FUNCTION $f_1(x)$

i	$f[x^{(i)}]$	$x_1^{(i)}$	$x_2^{(i)}$
0	404.0	–1.0	–1.0
1	19.97	0.2576	–0.3743
2	0.8654	0.0707	0.00067
3	0.318	0.452	0.1910
4	0.3048	0.448	0.199
5	0.2929	0.472	0.211
6	0.2828	0.4695	0.218
29	0.1752	0.5861	0.3373
30	0.1728	0.5846	0.3403
73	0.1081	0.6739	0.4499
74	0.1071	0.6729	0.4517

2-7.1.2 COST FUNCTION: $f_2(x) = (x_1 + 10x_2)^2 + 5(x_3 - x_4)^2 + (x_2 - 2x_3)^4 + 10(x_1 - x_4)^4$

Exact solution: $(0,0,0,0)$, $f_2(0) = 0$

TABLE 2-2

STEEPEST DESCENT METHOD – ITERATIVE DATA FOR COST FUNCTION $f_2(x)$

i	$f[x^{(i)}]$	$x_1^{(i)}$	$x_2^{(i)}$	$x_3^{(i)}$	$x_4^{(i)}$
0	122.0	1.0	1.0	1.0	1.0
1	16.43	0.9065	0.055	1.0	1.0
2	16.31	0.9019	0.023	0.9958	0.9581
5	16.03	0.8925	–0.0498	0.969	0.746
6	15.06	0.886	–0.0756	0.923	0.463
9	12.25	0.641	–0.063	0.698	–0.156
10	3.00	–1.048	0.0746	–0.1608	–0.9107
11	2.006	–1.039	0.1522	–0.258	–0.815
12	1.380	–1.043	0.078	–0.298	–0.752
13	1.188	–1.033	0.1127	–0.3276	–0.7067
14	1.047	–1.021	0.090	–0.362	–0.634
15	1.041	–1.015	0.0949	–0.369	–0.619
16	1.040	–1.012	0.0960	–0.370	–0.611
38	1.039	–1.008	0.0967	–0.373	–0.603
74	1.039	–1.008	0.0968	–0.373	–0.6019

2-7.1.3 COST FUNCTION: $f_3(x) = x_1^2 + 2x_2^2 + 2x_3^2 + 2x_1x_2 + 2x_2x_3$

Exact solution: $(0,0,0)$, $f_3(0) = 0$.

TABLE 2-3

STEEPEST DESCENT METHOD – ITERATIVE DATA FOR COST FUNCTION $f_3(x)$

i	$f[x^{(i)}]$	$x_1^{(i)}$	$x_2^{(i)}$	$x_3^{(i)}$
0	9.0	1.0	1.0	1.0
1	0.0714	0.2857	–0.0715	–0.0715
2	0.01311	0.1632	–0.153	0.0512
3	0.0088	0.1604	–0.114	0.065
4	0.00679	0.1245	–0.062	0.053
5	0.00243	0.078	–0.0625	0.0204
6	0.0018	0.073	–0.0476	0.02727
7	0.00063	0.0218	–0.00305	0.0133
8	0.00006	0.014	–0.00956	0.0035
9	0.00005	0.011	–0.00686	0.00485
10	0.00003	0.0041	–0.0036	0.0040

It should be noted that the Steepest Descent Method decreased the cost function rapidly on the first iteration but in the first two problems failed to converge to the minimum point That is typical behavior for this method, particularly in problems for which the cost function has a long sharp valley. It should be clear that blind use of the Method of Steepest Descent can yield poor results.

2-7.2 GENERALIZED NEWTON METHOD

2-7.2.1 COST FUNCTION: $f_1(x) = 100(x_2 - x_1^2)^2 + (1 - x_1)^2$

Exact solution: $(1,1)$, $f_1(1,1) = 0$

TABLE 2-4

GENERALIZED NEWTON METHOD – ITERATIVE DATA FOR COST FUNCTION $f_1(x)$

i	$f_1[x^{(i)}]$	$x_1^{(i)}$	$x_2^{(i)}$
0	404.0	−1.0	−1.0
1	3.981	−0.9950	0.9869
2	3.403	−0.7919	0.5832
3	2.588	−0.5248	0.2241
4	1.549	−0.1832	−0.5105
5	0.953	0.0887	−0.0271
6	0.473	0.3642	0.1063
7	0.203	0.5355	0.3347
8	0.0531	0.8020	0.6315
9	0.0042	0.9536	0.9049
10	0.0002	0.9900	0.9810
11	2×10^{-6}	1.0003	1.0007

2-7.2.2 COST FUNCTION: $f_2(x) = (x_1 + 10x_2)^2 + 5(x_3 - x_4)^2 + (x_2 - 2x_3)^4 + 10(x_1 - x_4)^4$

Exact solution: (0,0,0,0), $f_2(0) = 0$

TABLE 2-5

GENERALIZED NEWTON METHOD – ITERATIVE DATA FOR COST FUNCTION $f_2(x)$

i	$f_2[x^{(i)}]$	$x_1^{(i)}$	$x_2^{(i)}$	$x_3^{(i)}$	$x_4^{(i)}$
0	137.0	1.0	1.0	1.0	2.0*
1	2.137	−0.3368	0.0175	0.3396	0.3249
2	0.0496	−0.0640	0.0250	0.1060	0.1229
3	0.0025	−0.0591	0.0047	0.0627	0.0617
4	0.0007	−0.0236	0.0031	0.0263	0.0271
5	0.00001	−0.0146	0.0014	0.0161	0.0160
6	1×10^{-6}	−0.0070	0.0007	0.0078	0.0079

*Note: The trial starting point (1,1,1,1) was a singular point for $\nabla^2 f_2$ so an alternate starting point was chosen and the algorithm converged.

2-7.2.3 COST FUNCTION: $f_3(x) = x_1^2 + 2x_2^2 + 3x_3^2 + 2x_1x_2 + 2x_2x_3$

Exact solution: (0,0,0), $f_3(0) = 0$.

TABLE 2-6

GENERALIZED NEWTON METHOD – ITERATIVE DATA FOR COST FUNCTION $f_3(x)$

i	$f_3[x^{(i)}]$	$x_1^{(i)}$	$x_2^{(i)}$	$x_3^{(i)}$
0	9.000	1.0	1.0	1.0
1	2×10^{-5}	0.0015	0.0015	0.0015

These results indicate that the Generalized Newton Method is indeed very powerful. Even in the second cost function where the initial estimate caused a singularity in $\nabla^2 f_2$, a second starting point yielded good results. Similar behavior has been noted in the literature, so one can expect to get good results with this method. It must be remembered, however, that this method requires that second derivatives of the cost function be computed.

2-7.3 CONJUGATE GRADIENT METHOD

2-7.3.1 COST FUNCTION: $f_1(x) = 100(x_2 - x_1^2)^2 + (1 - x_1)^2$

Exact solution: (1,1), $f_1(1,1) = 0$.

TABLE 2-7

CONJUGATE GRADIENT METHOD—ITERATIVE DATA FOR COST FUNCTION $f_1(x)$

i	$f[x^{(i)}]$	$x_1^{(i)}$	$x_2^{(i)}$
0	404.0	−1.0	−1.0
	9649.	0.1143	0.0102
			0.0839
1	9649.	0.3258	0.0102
	22.19	0.5106	0.2360
2	22.19	0.5005	0.2482
	0.5033	0.6307	0.3820
3	0.5033	0.6244	0.3882
	0.2226	0.7267	0.5178
4	0.2226	0.7227	0.5212
	0.001637	0.9919	0.9827
8	0.001637	0.9842	0.9868
	0.000067	0.000013	0.999754
11	0.000067	0.999884	0.999768

2-7.3.2 COST FUNCTION: $f_2(x) = (x_1 + 10x_2)^2 + 5(x_3 - x_4)^2 + (x_2 - 2x_3)^4 + 10(x_1 - x_4)^4$

Exact solution: $(0,0,0,0)$, $f_2(0) = 0$.

TABLE 2-8

CONJUGATE GRADIENT METHOD—ITERATIVE DATA FOR COST FUNCTION $f_2(x)$

i	$f[x^{(i)}]$	$x_1^{(i)}$	$x_2^{(i)}$	$x_3^{(i)}$	$x_4^{(i)}$
0	122.0	1.0	1.0	1.0	1.0
	2925.27	0.9016	0.0346	0.9642	1.000
	2925.27	0.8637	0.07787	0.4158	0.9960
	36.55	0.8561	0.0101	0.3642	0.6573
1	192.2	0.8404	0.0563	0.3324	0.4438
	192.2	0.8180	0.0860	0.3185	0.4540
	12.11	0.7507	0.0685	0.2651	0.4787
	3.023	0.6410	0.0850	0.2533	0.4104
2	26.29	0.3404	0.0281	0.2192	0.2025
	26.29	0.3378	0.0370	0.2079	0.2057
	1.651	0.3331	0.0319	0.1748	0.2130
	0.1136	0.3159	0.0371	0.1173	0.1698
6	0.0531	0.0305	0.0029	0.0718	0.0717
	0.000751	0.0293	0.0031	0.0696	0.0714
	0.000751	0.0294	0.0028	0.0695	0.0713
	0.000556	0.0297	0.0029	0.0687	[illegible]
9	0.00101	0.03519	0.003519	0.02338	0.02503
	0.000512	0.035322	0.003498	0.023112	0.023119
	0.000512	0.035318	0.003530	0.023108	0.023119
	1×10^{-7}	0.035305	0.003522	0.023011	0.023115

2-7.3.3 COST FUNCTION: $f_3(x) = x_1^2 + 2x_2^2 + 2x_3^2 + 2x_1x_2 + 2x_2x_3$

Exact solution: $(0,0,0)$, $f_3(0) = 0$.

TABLE 2-9

CONJUGATE GRADIENT METHOD—ITERATIVE DATA FOR COST FUNCTION $f_3(x)$

i	$f_3[x^{(i)}]$	$x_1^{(i)}$	$x_2^{(i)}$	$x_3^{(i)}$
0	9.0	1.0	1.0	1.0
	0.1181	0.3829	−0.2340	0.0744
	0.0293	0.2571	−0.2285	0.1428
1	0	0	0	0

The numerical results presented here indicate that the Conjugate Gradient Method is very effective even for the Rosenbrock function $f_1(x)$. The method requires approximately the same amount of computation per step as the Steepest Descent Method but shows spectacularly improved performance.

It should be noted, however, that convergence slows as the minimum point is approached. In fact, as shown in Table 2-8, convergence to the required accuracy was not attained in one case.

2-7.4 FLETCHER-POWELL METHOD

2-7.4.1 COST FUNCTION: $f_1(x) = 100(x_2 - x_1^2)^2 + (1 - x_1)^2$

Exact solution: $(1,1)$, $f_1(1,1) = 0$.

TABLE 2-10

FLETCHER-POWELL METHOD—ITERATIVE DATA FOR COST FUNCTION $f_1(x)$

i	$f_1[x^{(i)}]$	$x_1^{(i)}$	$x_2^{(i)}$
0	404.0	−1.0	−1.0
1	19.97	0.2570	−0.3746
2	0.7839	0.1146	0.01249
3	0.7570	0.1422	0.005683
4	0.7424	0.1727	0.005740
5	0.5377	0.3378	0.08262
6	0.4013	0.3689	0.1416
7	0.2968	0.4815	0.2151
8	0.2524	0.5616	0.2909
9	0.03621	0.8286	0.6784
10	0.03216	0.8207	0.6733
11	0.02568	0.8536	0.7221
12	0.01162	0.9268	0.8511
13	0.00437	0.9342	0.8733
14	0.00106	0.9760	0.9504
15	8×10^{-6}	0.9982	0.9967

2-7.4.2 COST FUNCTION: $f_2(x) = (x_1 - 10x_2)^2 + 5(x_3 - x_4)^2 + (x_2 - 2x_3)^4 + 10(x_1 - x_4)^4$

Exact solution: $(0,0,0,0)$, $f_2(0) = 0$.

TABLE 2-11

FLETCHER-POWER METHOD—ITERATIVE DATA FOR COST FUNCTION $f_2(x)$

i	$f[x^{(i)}]$	$x_1^{(i)}$	$x_2^{(i)}$	$x_3^{(i)}$	$x_4^{(i)}$
0	122.0	1.0	1.0	1.0	1.0
1	14.4292	0.9017	0.03472	0.9642	1.0
2	2.3775	0.8630	−0.07820	0.4120	0.9960
3	0.6678	0.8430	−0.08740	0.3618	0.4986
4	0.3353	0.2087	−0.02560	0.3644	0.3305
5	0.05134	0.1117	0.006686	0.1883	0.1952
6	0.01059	0.07931	−0.009696	0.1537	0.1526
7	0.00067	0.02731	−0.0007003	0.06189	0.06276
8	0.00016	0.02164	−0.002344	0.05417	0.05409
9	8.3×10^{-6}	0.00267	−0.0000359	0.0191	0.0192
10	2.1×10^{-6}	0.00148	0.000163	0.0172	0.0172
11	10^{-7}	0.0057	0.00060	0.00341	0.00342

2-7.4.3 COST FUNCTION: $f_3(x) = x_1^2 + 2x_2^2 + 2x_3^2 + 2x_1x_2 + 2x_2x_3$

Exact solution: $(0,0,0)$, $f_3(0) = 0$.

TABLE 2-12

FLETCHER-POWELL METHOD – ITERATIVE DATA FOR COST FUNCTION $f_3(x)$

i	$f_3[x^{(i)}]$	$x_1^{(i)}$	$x_2^{(i)}$	$x_3^{(i)}$
0	9.0	1.0	1.0	1.0
1	0.05319	0.3830	−0.2340	0.07447
2	0.02857	0.2571	−0.2286	0.1429
3	3×10^{-13}	2×10^{-7}	-2×10^{-7}	-3×10^{-7}

The Fletcher-Powell Method requires slightly more computation than the Conjugate Gradient Method. However, its convergence properties are very good as the minimum point is approached, in contrast to the behavior of the Conjugate Gradient Method.

This method appears to have good properties in all ranges of the iterative process. It is more stable than the Generalized Newton Method in the early stages of computation and converges more rapidly than the Gradient and Conjugate Gradient Methods near the minimum point. In these respects it has the desirable properties of other methods without having their undesirable properties.

2-7.5 CONJUGATE DIRECTIONS WITHOUT DERIVATIVES

2-7.5.1 COST FUNCTION: $f_1(x) = 100(x_2 - x_1^2)^2 + (1 - x_1)^2$

Exact solution: $(1,1)$, $f_1(1,1) = 0$.

TABLE 2-13

CONJUGATE DIRECTIONS WITHOUT DERIVATIVES METHOD–ITERATIVE DATA FOR COST FUNCTION $f_1(x)$

i	$f_1[x^{(i)}]$	$x_1^{(i)}$	$x_2^{(i)}$
0	404.0	−1.0	−1.0
	100.1	0.0049	−1.0000
	0.9902	0.0049	0.0000
1	0.9902	0.0261	0.0211
	0.9485	0.0261	0.0007
	0.9402	0.0429	0.0174
2	0.7922	0.1287	−0.0016
	0.7022	0.1815	0.0509
	0.0172	0.2147	0.0436
3	0.3958	0.4058	0.1440
4	0.2895	0.4785	0.2422
5	0.2591	0.5308	0.3015
6	0.0770	0.7258	0.5225
7	0.0282	0.8564	0.7246
8	0.0125	0.8942	0.8033
9	0.0119	0.9116	0.8373
10	0.0116	0.9039	0.8218
11	0.0125	0.9469	0.9065
12	0.0042	1.0363	1.0792
13	0.0002	0.9886	0.9781
14	0.0002	1.0032	1.0079

2-7.5.2 COST FUNCTION: $f_2(x) = (x_1 - 10x_2) + 5(x_3 - x_4) + (x_2 - 2x_3) + 10(x_1 - x_4)$

Exact solution: $(0,0,0,0)$, $f_2(0) = 0$.

TABLE 2-14

CONJUGATE DIRECTIONS WITHOUT DERIVATIVES METHOD–ITERATIVE DATA FOR COST FUNCTION $f_2(x)$

i	$f_2[x^{(i)}]$	$x_1^{(i)}$	$x_2^{(i)}$	$x_3^{(i)}$	$x_4^{(i)}$
0	122.0	1.0	1.0	1.0	1.0
	109.1	0.2051	1.0000	1.0000	1.0000
	18.45	0.2051	0.1140	1.0000	1.0000
	7.667	0.2051	0.1140	0.4819	1.0000
1	2.371	0.2051	0.1140	0.4819	0.4284
	2.157	0.0469	0.1140	0.4819	0.4284
	1.075	0.0469	0.0127	0.4819	0.4284

TABLE 2-14 (Continued)

i	$f_2[x^{(i)}]$	$x_1^{(i)}$	$x_2^{(i)}$	$x_3^{(i)}$	$x_4^{(i)}$
	0.4421	0.0469	0.0127	0.2799	0.4284
2	0.1415	0.0469	0.0127	0.2799	0.2423
	0.1418	0.0510	0.0127	0.2799	0.2423
	0.1210	0.0510	−0.0015	0.2799	0.2423
	0.0498	0.0510	−0.0015	0.1875	0.2423
	0.0246	0.0510	−0.0015	0.1875	0.1749
3	0.0082	0.0536	−0.0104	0.1291	0.1324
4	0.0020	0.638	−0.0181	0.0794	0.0882
5	0.0018	0.1322	−0.0147	0.0892	0.0940
6	0.0010	0.0828	−0.0109	0.0580	0.0603
7	0.0005	0.0412	−0.0057	0.0377	0.0322
8	0.0000	0.0078	−0.007	0.0058	0.0050

2-7.5.3 COST FUNCTION: $f_3(x) = x_1^2 + 2x_2^2 + 2x_3^2 + 2x_1x_2 + 2x_2x_3$

Exact solution: $(0,0,0)$, $f_3(0) = 0$.

TABLE 2-15

CONJUGATE DIRECTIONS WITHOUT DERIVATIVES METHOD–ITERATIVE DATA FOR COST FUNCTION $f_3(x)$

i	$f_3[x^{(i)}]$	$x_1^{(i)}$	$x_2^{(i)}$	$x_3^{(i)}$
0	9.0	1.0	1.0	1.0
	5.000	−1.0000	1.0000	1.0000
	3.000	−1.0000	0.0000	1.0000
1	1.000	−1.0000	0.0000	0.0000
	0.000	0.0000	0.0000	0.0000

The Conjugate Directions Without Derivatives Method is not as efficient as some of the methods that require computation of derivatives. However, there are many problems in which computation of derivatives is either impossible or very difficult. In these problems, this method appears to be effective.

2-8 AN APPLICATION OF UNCONSTRAINED OPTIMIZATION TO STRUCTURAL ANALYSIS

As pointed out earlier in this chapter, optimal design problems are seldom unconstrained. There is, however, a large class of analysis problems which can be solved using unconstrained optimization methods. In Appendix B, energy principles which govern equilibrium, vibration, and stability of structures are given. The condition for equilibrium is particularly direct since it requires that, in problems for which the strain energy is quadratic, the equilibrium state, $\bar{x}$, minimizes V of Eq. B-18, Appendix B,

$$V = \frac{1}{2} x^T K x - x^T F . \tag{2-24}$$

Even in some problems which are nonlinear and the total potential energy is not quadratic, the minimum energy principle applies.

In view of the quadratic form of Eq. 2-24, conjugate direction methods are indicated. Even for nonquadratic energy expressions, methods for conjugate directions appear to be very efficient. For a much more detailed treatment of this class of equilibrium problems, see Ref. 12.

A second structural analysis problem for which unconstrained optimization methods hold even more promise is the eigenvalue problem. As shown in Appendix B, vibration and buckling problems reduce to eigenvalue problems of the kind

$$Ky = \lambda My . \tag{2-25}$$

In this problem, the smallest eigenvalue λ_1 of the Eq. 2-25 is sought. One method of solving this problem is to rewrite Eq. 2-25 as

$$K^{-1} My = \frac{1}{\lambda} y \tag{2-26}$$

In this form, an iterative technique such as the power (or iteration) method (Ref. 13, page 93) may be applied to obtain the largest eigenvalue of the matrix $K^{-1}M$ and hence, the smallest eigenvalue of the original problem. Even though the power method is efficient, this approach has the severe disadvantage of requiring that K^{-1} be computed.

A more promising approach to the above eigenvalue problem utilizes the Rayleigh quotient (Ref. 13, page 83), i.e., the smallest eigenvalue λ_1 of Eq. 2-25 is given by

$$\lambda_1 = \min_{y \neq 0} \frac{y^T K y}{y^T M y} . \tag{2-27}$$

If the vector y is normalized by fixing one of its elements, the resulting vector denoted $\tilde{y}$, then Eq. 2-27 reduces to

$$\lambda_1 = \min_{\tilde{y}} \frac{\tilde{y}^T K \tilde{y}}{\tilde{y}^T M \tilde{y}} . \tag{2-28}$$

The minimization Eq. 2-28 may now be solved by any of the methods of the present chapter. The method of conjugate directions has been recently applied to solve this class of problems (Refs. 14, 15). It is interesting to note that this exact approach to the eigenvalue problem was proposed by the inventor of conjugate direction methods, M. R. Hestenes, in 1955 (Ref. 16, page 93). The technique was apparently not used in engineering problems, however, until 1966.

Iterative methods of the kind outlined in this paragraph are particularly appropriate for iterative optimal design techniques. As discussed in Chapter 5, the most time consuming task in iterative design methods is the reanalysis of the system during each iteration, i.e., after the design variable is changed slightly, analysis for stresses, displacements, and eigenvalues must be done even though it is expected that these quantities will be very close to their values before the change in design variables. By using an iterative technique such as conjugate directions, the previous state may be used as an estimate to start the minimization algorithm. In this way, rapid convergence to the new state of the system is attained. This approach has been applied with good success (Ref. 15).

REFERENCES

1. C. Goffman, *Calculus of Several Variables*, Harper and Row, New York, 1965.
2. S. Perlis, *Theory of Matrices*, Addison-Wesley, Reading, Mass., 1952.
3. T. L. Saaty and J. Bram, *Nonlinear Mathematics*, McGraw-Hill, New York, 1964.
4. D. J. Wilde and C. S. Beightler, *Foundations of Optimization*, Prentice-Hall, Englewood Cliffs, New Jersey, 1967.
5. A. V. Fiacco and G. P. McCormick, *Nonlinear Programming Sequential Unconstrained Minimization Techniques*, John Wiley & Sons, New York, 1968.
6. J. C. Pearson, *On Variable Metric Methods of Minimization*, Technical Paper RAC-TP-302, Research Analysis Corporation, McLean, Virginia, February 1968.
7. R. K. Williamson, *Direct Methods of Parameter Optimization for Dynamical Systems*, Air Force Rept No. SAMSO-TR-69-432, July 15, 1969.
8. W. C. Davidon, *Variable Metric Method for Minimization*, Atomic Energy Commission R.E.D. Rept. ANL-5990, 1959.
9. R. Fletcher and M. J. D. Powell, "A Rapidly Convergent Descent Method for Minimization", *The Computer J.*, Vol. 6, p. 163, 1963.
10. M. J. Box, "A Comparison of Several Current Optimization Methods and the Use of Transformations in Constrained Problems", *The Computer J.*, Vol. 9, p. 67, 1966.
11. M. J. D. Powell, "An Effective Method for Finding the Minimum of Functions of Several Variables Without Use of Derivatives", *British Computer J.*, Vol. 7, pp. 155-162, 1964.
12. R. L. Fox and E. L. Stanton, "Developments in Structural Analysis by Direct Energy Minimization," *AIAA J.*, Vol. 6, No. 6, June 1968, pp. 1036-1042.
13. S. H. Crandall *Engineering Analysis*, McGraw-Hill, New York, 1956.
14. W. W. Bradbury and R. Fletcher, "New Iterative Methods for Solution of Eigenproblem", *Numerische Mathematik*, 9, 1966, pp. 259-267.
15. M. P. Kapoor, *Automated Optimum Design of Structures Under Dynamic Response Restrictions*, Report No. 28, Division of Solid Mechanics Structures and Mechanical Design, Case Western Reserve University, Cleveland, Ohio, January 1969
16. M. R. Hestenes, "Iterative Computational Methods", *Comm. Pure Appl. Math.*, Vol. 8, 1955, pp. 85-96.

CHAPTER 3

LINEAR PROGRAMMING

3-1 INTRODUCTION

In the preceding chapter a function $f(x)$, x in R^n, was minimized with no restrictions placed on the location of the design variable x. Problems in the real world seldom reduce to this form. In virtually all engineering design problems, requirements are placed on the object being designed, and these requirements are stated in terms of equations involving the design variable. More often, these requirements may be stated in terms of inequalities involving the design variable.

Examples of inequality constraints are abundant in all areas of engineering design. The following are examples:

1. Optimal structural design

a. Stress must be less than or equal to the yield strength of the material.

b. Buckling load must be greater than or equal to applied loads.

c. Deflection of the structure must not exceed specified limits.

d. Natural frequency must lie within an allowable range.

2. Optimal circuit design:

a. Voltage must remain within linear range of components.

b. Power consumption must be below a specified level.

c. Capacitance of a proposed capacitor must be within attainable limits.

3. Aerospace vehicle guidance:

a. Controller thrust must be within the capability of the thruster.

b. Total fuel consumption for a mission must be less than or equal to the vehicle's storage capacity.

c. Altitude must be greater than or equal to zero.

This list of typical inequality constraints could be expanded many-fold. It is clear then that the inequality constraint must play a central role in any unified theory of design.

The class of problem considered in this chapter is very restricted. Only linear functions are to be minimized subject to constraints which are linear in the design variables. In matrix notation this is, minimize

$$f(x) = C^T x \tag{3-1}$$

where C is an $n \times 1$ matrix of constants. The design variable x is required to satisfy

$$\left.\begin{aligned} Ax &\leq B \\ x &\geq 0 \end{aligned}\right\} \tag{3-2}$$

where A is an $m \times n$ matrix and B is an $m \times 1$ matrix. The inequality, Eq. 3-2, is taken as

$$\sum_{j=1}^{n} a_{ij} \cdot x_j < b_i, \; i = 1, \ldots, n .$$

i.e., when one vector is less than or equal to another vector, each of the components of the vectors must satisfy this relation.

Example 3-1: Consider the problem of minimizing

$$f(x) = x_1 + 2x_2 \tag{3-3}$$

subject to the constraints

$$\left.\begin{aligned} 2x_1 + x_2 &< 4 \\ x_1 &> 0 \\ x_2 &> 0. \end{aligned}\right\} \tag{3-4}$$

The constraints, Eq. 3-4, are satisfied at all points in the triangular region of Fig. 3-1. The lines passing through this region are lines of constant value of $f(x)$. It is clear that as the line is translated downward, the value of $f(x)$ decreases and that the lowest line that still contains points in the admissible region occurs for $x_1 + 2x_2 = 0$. Since this line intersects the admissible region only at (0,0), $f(x)$ takes on an absolute minimum at (0,0).

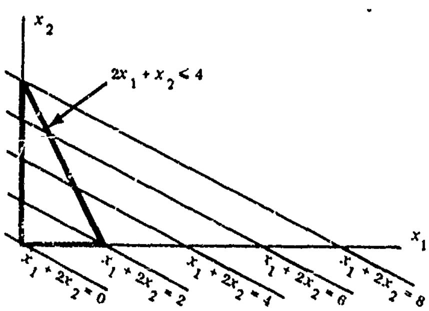


Figure 3-1. Graphical Solution of Example 3-1

As will be seen in the following paragraph, this is typical of linear programming problems.

Before proceeding to the next paragraph, it is worthwhile to discuss the applicability of linear programming. The theory of linear programming arose out of studies of economic activities. In economics it is often the case that behavior of an economic system is predictable only in a rather crude way, so frequently a linear relation among variables is as good a representation as can be expected.

In engineering design, however, it is very seldom that the behavior of an object or process can be described by linear expressions. One might be tempted, then, to completely ignore linear programming. Even though it is not directly applicable to most engineering design problems, however, linear programming is still a very powerful tool. First, even though the computational procedures of linear programming do not carry over to the real nonlinear world, many facets of the behavior of solutions are very similar in more general programming problems. The engineer who has mastered linear programming will go into the study of the much more complex nonlinear programming armed with a powerful tool – intuition. Further, the solution of many nonlinear problems can be reduced to the solution of a sequence of linear programming problems. For a review of some of these applications of linear programming methods see Ref. 1.

3-2 PROPERTIES OF LINEAR PROGRAMS

To formalize the discussion of the previous paragraph, the following definition is made.

Definition 3-1: The *linear programming* *problem* is the problem of determining that x in R^n which minimizes

$$\left.\begin{aligned} &B^T x && (3\text{-}5)\\ &\text{and which satisfies} &&\\ &Ax \geq C && (3\text{-}6)\\ &x \geq 0 && (3\text{-}7)\end{aligned}\right\}\text{LP}$$

where $C \neq 0$ is an $m \times 1$ matrix, A is an $m \times n$ matrix, B is an $n \times 1$ matrix and the symbolism $\leq$ ($\geq$) as applied to matrices means that the relation less than or equal to (greater than or equal to) holds for corresponding components of the matrices.

It should be pointed out that Eqs. 3-5 through 3-7 do not explicitly cover all linear optimization problems. For example, it may be required to maximize a linear objective function. Further, equality constraints may be imposed and negative values of the x_i may be allowed. However, all these variations on the linear programming problem may be put into the form of the problem previously considered. An objective function may be maximized by minimizing its negative, equality constraints are nothing more than a pair of inequality constraints (i.e., $y = 0$ if and only if $y \leq 0$ and $-y \leq 0$), and a negative x_i may always be written as the difference between two new non-negative variables. There is therefore, no loss of generality in considering only the problem expressed by Eqs. 3-5 through 3-7.

Definition 3-2: The *constraint set* for the linear programming problem of Def. 3-1 is the set of points in R^n which satisfy Eqs. 3-6 and 3-7.

The constraint set associated with a problem is just the set of design variables which describe an admissible object or process, i.e., one which performs the required service but is not necessarily optimal. In LP the constraint set is a polyhedron and, according to Def. 2-4, this constraint set is convex. Further, according to the same definition, the cost function $f(x)$ for LP is convex. If the constraint set is bounded and nonempty, it is necessarily also closed and all the hypotheses of Theorems 2-2 and 2-3 are satisfied. One then concludes that $f(x)$ has a strict absolute minimum in the constraint and that is has no other relative minima.

Further, if $f(x)$ had a minimum in the interior of the constraint set, the necessary condition of Theorem 2-1 implies

$$\frac{\partial f}{\partial x_i} = c_i = 0,\ i = 1, \ldots, n$$

which contradicts Def. 3-1 of LP. Therefore, $f(x)$ cannot have a minimum point in the interior of the constraint set but must take on its minimum at the boundary. Weyl has shown, in fact, that the solution must lie on one of the vertices of the polyhedral constraint set (Ref. 2).

In spite of this elementary theory, it is possible that a linear programming problem may not have a solution. This may happen for two reasons. First, the constraint set may be empty; and second, the constraint set may be unbounded and the cost function may be decreased without restriction. In order to facilitate discussion of these difficulties, Definition 3-3 is made.

Definition 3-3: If the constraint set of LP is nonempty (empty), the problem is called *feasible (infeasible)*. If the constraint set is unbounded and the cost function is not bounded below, then the problem is called *unbounded*.

The concept of the dual problem that will be used in constructing solutions of LP's will now be discussed. The dual problem will also play a major role in obtaining results for more general optimization problems.

Definition 3-4: The linear programming problem of maximizing

$$\left.\begin{aligned} & C^T y && (3\text{-}8) \\ & \text{for } y \text{ in } R^m \text{ satisfying} \\ & A^T y < B && (3\text{-}9) \\ & y > 0 \end{aligned}\right\} \text{LPD}$$

where the matrices A, B, and C are the same as in LP, and are called the *dual* of LP.

The results of Theorem 3-1 relating LP and LPD are proved in Ref. 3, page 41, and Ref. 4, page 118.

Theorem 3-1: Let x and y be in the constraint sets of LP and LPD, respectively. Then

1. $C^T y < B^T x$. (3-10)

2. If $C^T y = B^T x$ then x and y are the solutions of LP and LPD, respectively. (3-11)

3. If LP (LPD) is unbounded, then LPD (LP) is infeasible.

4. If LP (LPD) is feasible and LPD (LP) is infeasible, then LP (LPD) in unbounded.

These results are useful in constructing solutions of linear programming problems. They are also used in providing Theorem 3-2 that is central to linear programming theory.

Theorem 3-2: Let LP and LPD both be feasible. Then both have solutions $\bar{x}$ and $\bar{y}$, respectively, and $B^T\bar{x} = C^T\bar{y}$.

The proof of Theorem 3-2 is involved and does not yield a method of constructing solutions. It may be found in Ref. 3, page 44, or Ref. 4, page 118.

Since the solution of LP must lie on a vertex of the polyhedral constraint set, it suffices to check at most a finite number of points for the minimum. This procedure is followed in an organized way by beginning at any vertex of the constraint set. If the cost function cannot be decreased by moving along an edge of the polyhedron that intersects this vertex, then this vertex is the solution. If, however, the cost function decreases by moving along some edge, this policy is followed until a second vertex is reached and the cost function has been reduced. Since there are only a finite number of vertices and it is impossible to return to a previously occupied vertex, the process must terminate at the minimum over the constraint set.

In order to illustrate the argument presented in the preceding paragraph, consider Example 3-2.

Example 3-2: By moving along edges of the constraint set, solve the LP

minimize $f(x_1, x_2) = -2x_1 - x_2$

subject to

$$-x_1 > -1$$

$$-x_2 > -1$$

$$-2x_1 - 2x_2 > -3$$

$x_1, x_2 \geq 0.$

Solution: The polyhedral constraint set is shown in Fig. 3-2.

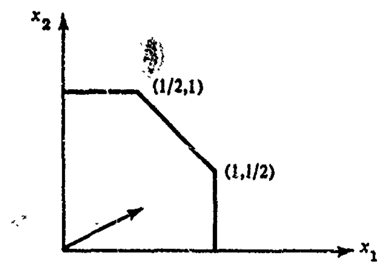


Figure 3-2. Polyhedral Constraint Set

The vector

$$-\alpha \nabla f^T(x_1, x_2) = \alpha \begin{bmatrix} 2 \\ 1 \end{bmatrix}$$

whose direction as shown in Fig 3-2 is the direction of steepest descent of $f(x)$. Starting at (0,0) a unit movement along the x_1-axis yields a change

$$df = \nabla f(0,0)dx = -2$$

and a unit movement along the x_2-axis yields a change

$$df = \nabla f(0,0)dx = -1$$

so both moves yield a decrease in $f(x)$. Choose the x_1-axis and move to the first vertex (1,0). The only movement possible is in the $+x_2$-direction from (1,0). A unit move in this direction yields

$$df = \nabla f(1,0)dx = -1$$

which decreases f. Move in this direction to the first vertex (1, 1/2).

The only move admissible is toward (1/2, 1). A unit move in this direction is obtained from

$$dx = \begin{bmatrix} -\frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} \end{bmatrix}$$

which causes a change in f,

$$df = \nabla f(1,1/2)dx = +\sqrt{2} - \frac{\sqrt{2}}{2} = \frac{\sqrt{2}}{2} > 0.$$

Therefore, f may not be decreased in moving from the vertex (1, 1/2) so this point is the solution of the problem.

The idea of moving from vertex to vertex is good for visualization but is poor for higher dimensional problems. The same idea, however, can be implemented algebraically. In order to obtain relations which will be required for solution of LP, define slack variables $u_1, \ldots, u_m$ so that

$$Ax - C = u \geq 0. \qquad (3\text{-}12)$$

The cost function of Eq. 3-5 will be denoted by the variable

$$w = B^T x. \qquad (3\text{-}13)$$

The problem LP now takes the form

$$\left.\begin{aligned} & Ax - C - u = 0 \\ & x \geq 0 \\ & u \geq 0 \\ & w = B^T x = \text{minimum} \end{aligned}\right\} \text{LP}'.$$

The solution of LP′ is the same as the solution of LP.

The information contained in Eqs. 3-12 and 3-13 is contained in the following matrix equation (called the simplex tableau):

$$\left[\begin{array}{cccc|c} a_{11} & a_{12} & \cdots & a_{1n} & -c_1 \\ a_{21} & a_{22} & \cdots & a_{2n} & -c_2 \\ \vdots & \vdots & & \vdots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} & -c_m \\ \hline b_1 & b_2 & \cdots & b_n & 0 \end{array}\right] \left[\begin{array}{c} x_1 \\ x_2 \\ \vdots \\ x_n \\ \hline 1 \end{array}\right] = \left[\begin{array}{c} u_1 \\ u_2 \\ \vdots \\ u_m \\ \hline w \end{array}\right] \qquad (3\text{-}14)$$

Eq. 3-14 may be viewed as $m + 1$ equations involving the variables $x_1, \ldots, x_n, u_1, \ldots, u_m, w$. At present Eq. 3-14 may be interpreted as determining $u_1, \ldots, u_m$, and w explicitly in terms of $x_1, \ldots, x_n$. It might be desirable to determine some other combination of $m + 1$ of the variables in terms of the remaining n. Except in singular cases, this is possible.

Assume now that $m + 1$ of the variables $s_1, \ldots, s_m$, and w have been determined explicitly in terms of the remaining n variables $r_1, \ldots, r_n$. Eq. 3-14 will then take the form

$$\left[\begin{array}{cccc|c} a'_{11} & a'_{12} & \cdots & a'_{1n} & -c'_1 \\ a'_{21} & a'_{22} & \cdots & a'_{2n} & -c'_2 \\ \vdots & \vdots & & \vdots & \vdots \\ a'_{m1} & a'_{m2} & \cdots & a'_{mn} & -c'_m \\ \hline b'_1 & b'_2 & & b'_n & \delta \end{array}\right] \left[\begin{array}{c} r_1 \\ r_2 \\ \vdots \\ r_n \\ \hline 1 \end{array}\right] = \left[\begin{array}{c} s_1 \\ s_2 \\ \vdots \\ s_m \\ \hline w \end{array}\right] \qquad (3\text{-}15)$$

where primes denote coefficients obtained when the original set of equations is solved for $s_1, \ldots, s_m$, and w.

The solution of LP will be constructed using a method which is based largely on Theorem 3-3.

Theorem 3-3: If in Eq. 3-15 $b'_i \geq 0, i = 1, \ldots, n$, and $-c'_j \geq 0$, $j = 1, \ldots, m$, then the solution of LP is

$$r_i = 0, i = 1, \ldots, n$$

$$s_j = -c_j, j = 1, \ldots, m$$

$$w = \delta.$$

It is clear from this theorem that any method of choosing the variables s_i and r_j which will terminate with non-negative entries in the last row and column, except perhaps for δ, will serve as a method of solving LP. Before developing such a method, several definitions will be helpful.

Definition 3-5: In Eq. 3-15, the variables s_j, $j = 1, \ldots, m$, are called *basic variables*, while the variables r_i, $i = 1, \ldots, n$ are called *nonbasic variables.*

Definition 3-6: The set of variables $s_1, \ldots, s_m, r_1, \ldots, r_n$ will be called a *basic point*. If $c_j < 0$, $j = 1, \ldots, m$, in Eq. 3-15, then the basic point will be called a *basic feasible point.*

A certain geometric interpretation may now be given for the nonbasic variables. In LP′ it is clear that the boundary of the constraint set of LP is obtained by setting various combinations of the variables x_i, $i = 1, \ldots, n$ and u_j, $j = 1, \ldots, m$, equal to zero. In the space R^n of the design variable x, a vertex of the polyhedral constraint set is obtained by having n equality constraints among the x_i, $i = 1, \ldots, n$, enforced. By the discussion, this occurs when $r_i = 0, i = 1, \ldots, n$. An edge of this polyhedron is a line in R^n obtained by setting $r_i = 0$ for $n - 1$ indices i. From Def. 3-6 and

Eq. 3-15, it is clear that a basic feasible point corresponds to a vertex of the polyhedral set. This is true since setting the nonbasic variables of the basic feasible point equal to zero yields admissible basic variables. Further, two vertices lie on the same edge of the constraint set if they have $n - 1$ of their nonbasic variables in common.

The process for interchanging the roles of a basic and a nonbasic variable thus becomes the central tool for methods based on Theorem 3-3. Suppose it is desired to make s_i a nonbasic variable and r_j a basic variable. If $a'_{ij} \neq 0$ then the ith equation from Eq. 3-15,

$$a'_{i1} r_1 + \ldots + a'_{ij} r_j + \ldots + a'_{in} r_n - c_i = s_i$$

may be solved for r_j to obtain

$$\begin{aligned} r_j = \frac{c'_i}{a'_{ij}} - \frac{a'_{i1}}{a'_{ij}} r_1 - \ldots - \frac{a'_{ij-1}}{a'_{ij}} r_{j-1} \\ + \frac{s_i}{a'_{ij}} - \frac{a'_{ij+1}}{a'_{ij}} r_{j+1} \\ - \ldots - \frac{a'_{in}}{a'_{ij}} r_n . \end{aligned} \tag{3-16}$$

Using this expression for r_j, r_j may be eliminated from the left sides of the remaining equations in Eq. 3-15. For $k \neq i$ this yields

$$\begin{aligned} &\left[a'_{k1} - \frac{a'_{i1} a'_{kj}}{a'_{ij}} \right] r_1 \\ &+ \ldots + \left[a'_{kj-1} - \frac{a'_{ij-1} a'_{kj}}{a'_{ij}} \right] r_{j-1} \\ &+ \frac{a'_{kj}}{a'_{ij}} s_i + \left[a'_{kj+1} - \frac{a'_{ij+1} a'_{kj}}{a'_{ij}} \right] r_{j+1} \\ &+ \ldots + \left[a'_{kn} - \frac{a'_{in} a'_{kj}}{a'_{ij}} \right] r_n \\ &- \left[c'_k - \frac{c'_i a'_{kj}}{a'_{ij}} \right] = s_k . \end{aligned} \tag{3-17}$$

It is thus clear how the coefficients in Eq. 3-15 change as the roles of a pair of variables are interchanged. This process may be described concisely in the language of Definition 3-7.

Definition 3-7: The entry $a'_{ij} \neq 0$, preceding Eq. 3-16, is called the *pivot* of the transformation. The transformation itself is called a *pivot step*.

The effect of the pivot step on the coefficient matrix of Eq. 3-15 may be illustrated easily by the diagram

$$\begin{bmatrix} p & \alpha \\ \beta & \gamma \end{bmatrix} \rightarrow \begin{bmatrix} \frac{1}{p} & -\frac{\alpha}{p} \\ \frac{\beta}{p} & \gamma - \frac{\alpha\beta}{p} \end{bmatrix} . \tag{3-18}$$

The diagram shown by Eq. 3-18 simply relates that in the coefficient matrix of Eq. 3-15 the following changes occur. The pivot is replaced by its inverse. All other elements in the same row as the pivot are multiplied by the negative inverse of the pivot. All other elements in the same column as the pivot are multiplied by the inverse of the pivot. All other elements in the matrix are decreased by the product of the element in their column and the row of the pivot, the element in their row and the column of the pivot and the inverse of the pivot.

Example 3-3:

Given $\begin{bmatrix} 2 & 1 & -4 \\ 3 & 6 & 1 \\ -5 & 3 & 2 \end{bmatrix} \begin{bmatrix} r_1 \\ r_2 \\ 1 \end{bmatrix} = \begin{bmatrix} s_1 \\ s_2 \\ w \end{bmatrix}$

interchange the role of r_1 and s_2.

Solution: The new matrix relation is

$$\begin{bmatrix} 2/3 & -3 & -14/3 \\ 1/3 & -2 & -1/3 \\ -5/3 & 13 & 11/3 \end{bmatrix} \begin{bmatrix} s_2 \\ r_2 \\ 1 \end{bmatrix} = \begin{bmatrix} s_1 \\ r_1 \\ w \end{bmatrix}.$$

It is shown in Ref. 3, page 53, that this pivoting transformation preserves the dual linear programming problem.

The pivoting transformation is an organized tool which allows one to interchange basic and nonbasic variables. It remains only to obtain an algorithm which uses this tool and Theorem 3-3 to construct the solution of LP.

3-3 THE SIMPLEX ALGORITHM

As was shown in par. 3-2, the solution of the linear programming problem may be reduced to the choice of pivot points. The algorithm presented here will have two phases. The first phase will consist of an algorithm for obtaining a basic feasible point. The second phase will operate only with basic feasible points and will successively reduce the cost function until the hypotheses of Theorem 3-3 are satisfied.

For convenience in the discussion which follows, it is assumed that the choice of basic and nonbasic variables has been made at a given stage of the solution process and the primes of Eq. 3-15 are dropped, i.e.,

$$\left[\begin{array}{ccc|} a_{11} & \cdots & a_{1n} & -c_1 \\ \vdots & & \vdots & \vdots \\ a_{m1} & \cdots & a_{mn} & -c_m \\ \hline b_1 & \cdots & b_n & \delta \end{array}\right] \begin{bmatrix} r_1 \\ \vdots \\ r_n \\ \hline 1 \end{bmatrix} = \begin{bmatrix} s_1 \\ \vdots \\ s_m \\ \hline w \end{bmatrix} \qquad (3\text{-}19)$$

Primes will now be used to denote the coefficients that result from a pivot step applied to Eq. 3-19. These new coefficients are determined by applying Eq. 3-18.

3-3.1 DETERMINATION OF A BASIC FEASIBLE POINT

If some elements in the right-hand column of the matrix of Eq. 3-19 (other than δ) are negative, then the present choice of variables is not a basic feasible point. Let $-c_k$ be the negative entry nearest the bottom of the column (again excluding δ). Since when $r_j = 0$, $j = 1, \ldots, n$, $s_k = -c_k < 0$, if there are admissible points in the constraint set of LP, then it must be possible to increase s_k by increasing some r_j from zero; i.e., there must be some positive a_{kj}. Choose j_0 so that $a_{kj_0} > 0$. This fixes the column index of the pivot.

To find an admissible row index i_0, consider first that after the pivot step

$$-c'_{i_0} = \frac{c_{i_0}}{a_{i_0 j_0}}.$$

It is clear then that candidates for the pivot $a_{i_0 j_0}$ must be limited to indices i for which

$$\frac{c_i}{a_{ij_0}} > 0. \qquad (3\text{-}20)$$

With this restriction in mind, consider the values of c_i' after the pivot step with $i \neq i_0$.

These are

$$-c_i' = -c_i + \frac{c_{i_0} a_{ij_0}}{a_{i_0 j_0}} \tag{3-21}$$

In order to insure $-c_i' > 0,\ i > k$, it is required that

$$-c_i + \frac{c_{i_0} a_{ij_0}}{a_{i_0 j_0}} > 0, i > k, i \neq i_0 . \tag{3-22}$$

If $a_{ij_0} > 0$ this clearly holds. If $a_{ij_0} < 0$, however, the requirement, Eq. 3-22, may be rewritten as

$$\frac{c_i}{a_{ij_0}} > \frac{c_{i_0}}{a_{i_0 j_0}}, i > k, i \neq i_0 . \tag{3-23}$$

Further, for $i = k$.

$$-c_k' = -c_k + \frac{c_{i_0} a_{kj_0}}{a_{i_0 j_0}} > -c_k \tag{3-24}$$

since $a_{kj_0} > 0$.

Inequalities, Eqs. 3-23 and 3-24, show that if i_0 is chosen so that

$$\frac{c_{i_0}}{a_{i_0 j_0}} = \min_{i > k} \left(\frac{c_i}{a_{ij_0}} \,\middle|\, \frac{c_i}{a_{ij_0}} > 0 \right) \tag{3-25}$$

then $-c_i' > 0$, $i > k$ and $-c_k' > -c_k$. If $-c_k$ is still negative, the process may be repeated. Otherwise choose the next entry above $-c_k$ which is negative and repeat the process.

If all the c_i, $i > k$ are nonzero, only a finite number of basic points are possible since the process is monotone (nonrepeating). If there exists a point with $-c_i > 0$, $i = 1, \ldots, m$, this process must find it. The degenerate case in which some $c_i = 0$, $i > k$ is discussed later.

The process described may be given quite simply as the iterative *Algorithm LP-A:*

Step 1. Choose $-c_k$ as the lowest negative entry (with the exception of δ) in the right-hand column of the coefficient matrix of Eq. 3-19.

Step 2. Choose any positive element a_{kj_0} in [illegible] row of the matrix of Eq. [illegible]

Step 3 Choose i_0 as in Eq. 3-25.

Step [illegible] Perform the pivot step with pivot c_i [illegible]

Step 5. If any $-$ [illegible] 0, $i = 1, \ldots, k$, choose that one with largest index i and return to Step 1. If $-c_i > 0$, $i = 1, \ldots, m$, then a basic feasible solution has been found and the process may be terminated.

3-3.2 SOLUTION OF LP

In par. 3-3.1 an algorithm is given for finding a basic feasible point. Once this has been accomplished, the object is to find a second algorithm which successively reduces w.

Since by Eq. 3-19, $w = b_1 r_1 + \ldots + b_n r_n + \delta$, it is clear that if $b_{j_0} < 0$ for some $j = j_0$ then w may be reduced by increasing r_{j_0} from zero. If a pivot step is performed which makes r_{j_0} a basic variable then w will be decreased The choice of the basic variable s_{i_0} which is to be made nonbasic must be made in such a way that the point obtained after the pivot step is still a basic feasible point, i.e., so that $-c_i > 0$, $i = 1, \ldots, m$. However, this is precisely the restriction which led to the choice of i_0 in par. 3-3.1. Therefore, the same procedure for choosing i_0 may be employed here.

Since $w' = w - c_{i_0} b_{j_0} / a_{i_0 j_0}$, the pivot step determined here guarantees $w' < w$ provided all $-c_i > 0$, $i = 1, ..., m$. In this case, therefore, only a finite number of pivot steps may be made, and the process must terminate at the solution of the linear programming problem. Termination occurs when $b'_j > 0, j = 1, ..., n$. Theorem 3-3 shows that this is the solution of the linear programming problem. The degenerate case where some $c_i = 0$ will be discussed par. 3-3.3.

This process is given explicitly in *Algorithm LP-B*

Step 1 Choose any negative entry (except 5) b_{j_0} in the bottom row of the coefficient matrix of Eq. 3-19.

Step 2. Choose i_0 according to Eq. 3-25 with $k = 1$.

Step 3. Perform the pivot step with pivot $a_{i_0 j_0}$.

Step 4. If any $b_j < 0$, $j = 1, ..., n$, choose one $b_{j_0} < 0$ and return to Step 1. If $b_j \geqslant 0$, $j = 1, ..., n$, then the solution of LP has been found.

3-3.3 THE DEGENERATE CASE

In both pars. 3-3.1 and 3-3.2 the computational algorithms could have problems if some $c_i = 0$. This situation is called degenerate since when n constraints are made equalities by putting $r_j = 0, j = 1, ..., n$, one has $s_i = c_i = 0$ which means that still another constraint is an equality. The degeneracy arises from the fact that in LP the n dimensional design variable $x = (x_1, ..., x_n)$ satisfies $n+1$ linear equalities. Therefore, the $n + 1$ equations are not linearly independent.

Viewed geometrically, the difficulty occurs because the path which successive basic points follow on the polygonal constraint boundary may form a closed loop. To prevent this behavior with only a small error in the final solution an entry, $-c_i$, which is zero, is replaced by an arbitrarily small parameter $\epsilon > 0$. The problem is not degenerate any longer and cycling cannot occur. Therefore, the altered problem will proceed toward the solution.

Example 3-4: Use the simplex algorithm to solve the LP

minimize $2x_1 + 9x_2 + x_3$

subject to

$$x_1 + 4x_2 + 2x_3 > 5$$

$$3x_1 + x_2 + 2x_3 > 4$$

$$x_1 > 0$$

$$x_2 > 0$$

$$x_3 > 0 .$$

First, LP′ is:

minimize w where

$$\begin{bmatrix} 1 & 4 & 2 & -5 \\ 3 & 1 & 2 & -4 \\ 2 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ 1 \end{bmatrix} = \begin{bmatrix} u_1 \\ u_2 \\ w \end{bmatrix}$$

subject to

$$x_i > 0, \quad i = 1, 2, 3, \quad u_j > 0, \quad j = 1, 2$$

For the first pivot step in algorithm LP-A, $k = 2$. Choose $j_0 = 1$ since $a_{21} = 3$ is the largest element in the second row. $i_0 = 2$ is the only choice available in Eq. 3-25 and

$$\frac{c_2}{a_{22}} = \frac{4}{1} = 4 > 0 .$$

The pivot is $a_{22} = 1$. This pivot step interchanges u_2 and x_2. The result is

$$\begin{bmatrix} -11 & 4 & -6 & 11 \\ -3 & 1 & -2 & 4 \\ -25 & 9 & -17 & 36 \end{bmatrix} \begin{bmatrix} x_1 \\ u_2 \\ x_3 \\ 1 \end{bmatrix} = \begin{bmatrix} u_1 \\ x_2 \\ w \end{bmatrix}.$$

Note that this basic point is already a basic feasible point so that the process now transfers to algorithm LP-B. Since b'_i is most negative, choose $j_0 = 1$. Now,

$$\frac{c_1}{a_{11}} = 1 \quad \frac{c_2}{a_{21}} = \frac{4}{3},$$

so $i_0 = 1$. The pivot is then $a_{11} = -11$. The result of a pivot step is to interchange x_1 and u_1. This results in the basic feasible point

$$\begin{bmatrix} -1/11 & 4/11 & -6/11 & 1 \\ 3/11 & -1/11 & -4/11 & 1 \\ 25/11 & -1/11 & -37/11 & 11 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ x_3 \\ 1 \end{bmatrix} = \begin{bmatrix} x_1 \\ x_2 \\ w \end{bmatrix}.$$

Choose $j_0 = 3$.

$$\frac{c_1}{a_{13}} = 11/6, \frac{c_2}{a_{23}} = 11/4,$$

so $i_0 = 1$. The pivot is $a_{22} = -6/11$ and a pivot step leads to

$$\begin{bmatrix} -1/6 & 2/3 & -11/6 & 11/6 \\ 1/3 & -1/3 & 2/3 & 1/3 \\ 17/6 & -7/3 & 37/6 & 29/6 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ x_1 \\ 1 \end{bmatrix} = \begin{bmatrix} x_3 \\ x_2 \\ w \end{bmatrix}.$$

Put $j_0 = 2$,

$$\frac{c_1}{a_{12}} = -11/4, \frac{c_2}{a_{22}} = 1,$$

so $i_0 = 2$ and $a_{22} = -1/3$ is the pivot. A pivot step yields

$$\begin{bmatrix} 1/2 & -2 & -1/2 & 5/2 \\ 1 & -3 & 2 & 1 \\ 1/2 & 7 & 3/2 & 5/2 \end{bmatrix} \begin{bmatrix} u_1 \\ x_2 \\ x_1 \\ 1 \end{bmatrix} = \begin{bmatrix} x_3 \\ u_2 \\ w \end{bmatrix}$$

Since this is a basic feasible point and the first three elements in the third row are positive, then the solution is immediate. The nonbasic variables are zero,

$$u_1 = x_1 = x_2 = 0$$

and the basic variables take on the value

$$x_3 = 5/2, \ u_2 = 1, \text{ and } w = 5/2.$$

Therefore, the solution to the original LP is

$$x_1 = 0$$

$$x_2 = 0$$

$$x_3 = 5/2.$$

The minimum value of $f(x)$ attained is 5/2.

3-4 MINIMUM WEIGHT TRUSS DESIGN

As will become apparent in subsequent chapters, most optimal design problems are nonlinear. Even the problems considered in this paragraph appear at first glance to be nonlinear. However, it is shown that the problem can actually be solved as a linear program. This will not be the case in general. The class of problems and their solutions that are discussed in this paragraph are taken from an outstanding paper by Dorn, Gomory, and Greenberg (Ref. 5). Similar results have been reported more recently (Ref. 6).

The problem treated here is minimum weight design of plane trusses with constraints on stress. The initial restrictions on the truss include only the location of joints in the truss. The loads to be supported by the truss are applied at joints. A member with non-negative cross-sectional area is allowed to connect each pair of joints. If there are μ joints, there may be $\mu(\mu - 1)/2$ members in the truss. In general, then, statically indeterminate trusses are allowed.

Let A_j, $j = 1, \ldots, n$, denote the cross-sectional area of jth member and S_j the load in that member due to the external loads applied to the truss; $S_j > 0$ denotes tension. If $m = 2\mu$, then equilibrium of the joints of the truss is specified by the equations

$$\sum_{j=1}^{n} a_{ij} S_j = F_i, \quad i = 1, \ldots, m \tag{3-26}$$

where F_i are components of applied forces at the joints, and a_{ij} are direction cosines of the elements of the structure intersecting the ith joint. All a_{ij} are zero if the jth element does not intersect the point of application of F_i. In order to satisfy three equilibrium equations for the applied loads (including reactions at supports), it is assumed there are $m^* = m - 3$ linearly independent equations in Eq. 3-26.

If σ is the maximum allowable stress (both tensile and compressive) for the material from which the truss is constructed, then stress constraints are

$$|S_j| < \sigma A_j . \tag{3-27}$$

Further, if ρ is the weight density of the structural material, the total weight W of the truss which is to be minimized is

$$W = \rho \sum_{j=1}^{n} A_j \ell_j \tag{3-28}$$

where ℓ_j is the length of the jth member.

The problem of minimizing W of Eq. 3-28 subject to the constraints of Eqs. 3-26 and 3-27 is not the complete truss design problem. In addition to the equilibrium conditions of Eq. 3-26, a set of compatibility conditions between displacements of the joints must be satisfied. These compatibility conditions will be nonlinear in the variables S_j and A_j. In its complete formulation, then, the truss design problem is not a linear programming problem. It will be shown, however, that if the compatibility conditions are ignored and the problem described by Eqs. 3-26, 3-27, and 3-28 is solved, its solution satisfies the compatibility conditions and is, therefore, the solution of the truss design problem.

Recalling that compatibility relations are being ignored, it is required that

$$|S_j| = \sigma A_j \quad j = 1, \ldots, n. \tag{3-29}$$

This is true since if $|S_j| < \sigma A_j$ for some j, then A_j could be reduced with an accompanying reduction in W. The constraint, Eq. 3-27, is therefore replaced by Eq. 3-29. The reader should note that this argument would not be valid if compatibility conditions were being enforced, since a reduction in some A_j may result in a violation of a constraint not involving A_j explicitly.

Since by Eq. 3-29, $A_j = \frac{1}{\sigma}|S_j|$, the optimization problem is now to minimize

$$W = \frac{\rho}{\sigma} \sum_{i=1}^{n} |S_j| \ell_j$$

subject to Eq. 3-26. In order to treat this problem as a linear programming problem, define

$$S_j^+ = \begin{cases} S_j, & \text{if } S_j \geq 0 \\ 0, & \text{if } S_j < 0 \end{cases}$$

$$S_j^- = \begin{cases} 0, & \text{if } S_j > 0 \\ -S_j, & \text{if } S_j < 0 \end{cases}$$

Now,

$$S_j = S_j^+ - S_j^-$$

and

$$|S_j| = S_j^+ + S_j^-$$

Denote

$$x^T = (S_1^+, \ldots S_n^+, S_1^-, \ldots, S_n^-)$$

$$C^T = (F_1, \ldots, F_m),$$

$$A = (a_{ij} \mid -a_{ij})_{m \times 2n}$$

and

$$B^T = \frac{\rho}{\sigma} (\ell_1, \ldots, \ell_n, \ell_1 \ldots, \ell_n).$$

In this notation, the problem is of the form LP′, namely, find x to minimize

$$B^T x \tag{3-30}$$

subject to

$$Ax - C = 0 \tag{3-31}$$

$$x > 0. \tag{3-32}$$

This linear programming problem may now be solved by the simplex method. Before the solution of the linear programming problem can be taken as the solution of the truss design problem, however, it must be shown that it satisfies the compatibility conditions. It is clear that if the truss specified by the linear programming problem is statically determinate, it satisfies the compatibility conditions trivially (i.e., there are no compatibility conditions). For the analysis here, statically determinate is taken to mean that the member forces S_j are uniquely determined by the given loads and the equilibrium conditions of Eq. 3-26.

As pointed out in Ref. 5, page 32, there will be m^* possibility nonzero components of x (basic variables) in the solution, corresponding to linearly independent columns of the matrix A; i.e., only m^* of the S_j will possibly be nonzero. According to Eq. 3-27, then, only m^* of the areas may be nonzero. Further, since the rank of A is m^*, the member forces are uniquely determined. The resulting truss is, therefore, statically determinate and hence is the solution of the original truss design problem.

It is pointed out (Ref. 5) that the simplex method for solving many member truss design problems is relatively time-consuming. It is proposed that the method be refined for this class of problems to obtain a practical method of solving engineering design problems. Several examples are solved in considerable detail in Ref. 5; the results of one of these problems will be discussed here.

A bridge truss is to be designed to span two points, 1 and 13 of Fig. 3-3. Three vertical levels of joints are allowed with five horizontal sets, a total of 15 points, as shown in Fig. 3-3. In the general case there could be 15(14)/2 = 105 members in the truss. Loads on the floor of the truss are shown in Fig. 3-3.

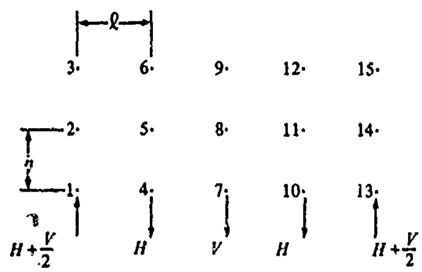


Figure 3-3. Admissible Joints for Bridge Truss

In the solution presented in Ref. 5, it is assumed that the truss is symmetric about the line of joints 7-8-9. This assumption reduces the number of variables to 57. Further, due to the assumed symmetry, there are only 14 independent equilibrium conditions. Therefore, there will be only 14 members which can be nonzero in the optimum truss. In the solution presented in Ref. 5 the problem is made nondimensional by defining $\alpha = h/\ell$ and $\beta = H/V$ where h and ℓ are the vertical and horizontal spacing, respectively, and H and V are applied loads shown in Fig. 3-3.

The solution presented in Ref. 5, page 45, for a fixed value of $\beta(\beta = 1)$ shows that there are three subintervals of values of α on each of which the truss has a constant geometrical form. For different values of α within a given subinterval, the member sizes are different. A plot of W vs α and the forms of optimal trusses are shown in Fig. 3-4.

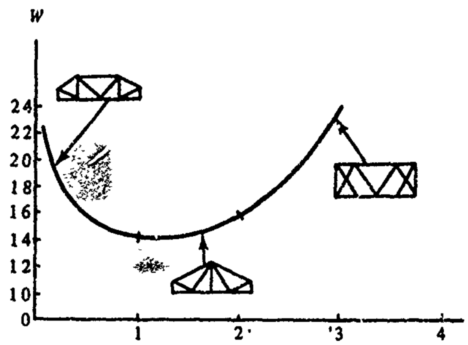


Figure 3-4. Optimum Bridge Trusses

The discussion here only touches on the highlights of the very complete treatment of the truss design problem in Ref. 5. The interested reader is encouraged to study this outstanding article in detail.

Before leaving the truss design problem, a point of interest in the present results and in the results obtained in future chapters may be noted. In Fig. 3-4 it is clear that at two values of α the form of the optimal truss changes form drastically. Still, even though the topology of the structure is not continuous in α, the weight apparently is a continuous function of α. The same sort of behavior occurs in a beam design problem with constraints on deflection which is discussed in par. 7-4. These problems might lead one to suspect that there is some basic mathematical structure of the optimal structural design problem that has not been uncovered.

3-5 AN APPLICATION OF LINEAR PROGRAMMING TO ANALYSIS

A major application of linear programming in engineering design is, oddly enough, in nonlinear programming. It is seldom that a realistic engineering design problem can be formulated as an LP. Realistic problems are generally nonlinear when considered as a function of both state and design variables. Several techniques of solution of nonlinear programming problems are based on approximation of the nonlinear problem by a linear one, at least locally. These methods then require that the approximating LP be solved. This subject will be deferred until a discussion of the general theory of nonlinear programming has been given.

A second application of linear programming which is of concern to the engineer is in the solution of linear boundary-value problems that arise in such fields as continuum mechanics. It should be emphasized here that this application is not of an optimal design nature, but rather falls in the field of engineering analysis.

One of the important methods of solving linear boundary-value problems is to approximate the solution by a linear combination of known functions. The question arises, "How should the coefficients be chosen so as to obtain the 'best' approximation to the true solution?" "Best" may be defined in many ways. A relatively new concept of "best" will be discussed in this paragraph.

The general linear boundary-value problem may be stated in operator notation as

$$L[z] = Q(x), \quad x \text{ in } \Omega \tag{3-33}$$

$$B[z] = q(x), \quad x \text{ on } \Gamma \tag{3-34}$$

where Ω is the domain of the independent variable $x \in R^n$ and Γ is its boundary. The dependent variable is a vector function of x, $z(x)$ in R^m. In the case of ordinary differential equations on $x_1 < x < x_2$

$$L[z] = \sum_{i=0}^{m} a_i(x) \frac{d^i z}{dx^i} \tag{3-35}$$

and the boundary operator is

$$B[z] = Az(x_1) + Bz(x_2). \tag{3-36}$$

In the case of partial differential equations,

$$L[z] = \sum_{|\alpha| < m} a_\alpha(x) \frac{\partial^{|\alpha|} z}{\partial x_1^{\alpha_1} \ldots \partial x_n^{\alpha_n}} \tag{3-37}$$

and the boundary operator is

$$B[z] = A(x)z(x), x \text{ on } \Gamma. \tag{3-38}$$

The method to be discussed treats both the partial and ordinary differential equations in the same way. Let $\phi_j(x)$, $j = 1, \ldots, k$ satisfy the homogeneous differential equation

$$L[\phi_j] = 0, \text{ in } \Omega. \tag{3-39}$$

Further, let $\phi_0(x)$ be found such that

$$L[\phi_0] = Q(x), \text{ in } \Omega. \tag{3-40}$$

Since the operator L is linear, the new function

$$\bar{z} = \phi_0 + \sum_{j=1}^{k} c_j \phi_j(x) \tag{3-41}$$

satisfies the differential Eq. 3-33 regardless of the value of the constants c_j. The object is now to find these constants so that $\bar{z}$ satisfies the boundary conditions of Eq. 3-34 as closely as possible.

Define

$$\| B[\bar{z}] - q \| = \max_i |B_i[z] - q_i| . \quad (3\text{-}42)$$

In this notation, $\bar{z}$ will be the solution of the boundary-value problem if and only if

$$\| B[\bar{z}] - q(x) \| = 0 \quad (3\text{-}43)$$

for all points x on Γ.

The method to be treated here attempts to minimize the error in Eq. 3-43 at a large number of points x^ℓ, $\ell = 1, \ldots, L$, on Γ. Define

$$\gamma = \max_\ell \| B[\bar{z}(x^\ell)] - q(x^\ell) \| . \quad (3\text{-}44)$$

The object now is to choose the constants c_j so as to minimize γ. To see that this is a linear programming problem, note that Eq. 3-44 is equivalent to

$$B_i[\bar{z}(x^\ell)] - q_i(x^\ell) < \gamma \quad (3\text{-}45)$$

and

$$- B_i[\bar{z}(x^\ell)] + q_i(x^\ell) < \gamma \quad (3\text{-}46)$$

for all i and ℓ.

Note that Eqs. 3-45 and 3-46 are linear in the c_j and γ. Since the c_j may be either positive or negative, it is necessary to define new constants $c_j^+ > 0$ and $c_j^- > 0$ such that

$$c_j = c_j^+ - c_j^- . \quad (3\text{-}47)$$

Now, the problem of choosing γ, c_j^+, c_j^- (all non-negative) which satisfy Eqs. 3-45 and 3-46 and which minimize γ is clearly a LP. Further, it is just a restatement of the best approximation criterion of Eq. 3-44.

Example 3-4: Obtain an approximate solution of

$$\Delta z = \frac{\partial^2 z}{\partial x_1^2} + \frac{\partial^2 z}{\partial x_2^2} + \frac{1}{x_2}\frac{\partial z}{\partial x_2} = 1 \quad (3\text{-}48)$$

in $\Omega = \{(x_1, x_2) \mid |x_1| < 1, 0 < x_2 < 1\}$ with

$$z + \frac{\partial z}{\partial n} = 0 \text{ on } \Gamma = \{(x_1, x_2) \mid |x_1| = 1,$$

$$x_2 = 0 \text{ or } x_2 = 1\} \quad (3\text{-}49)$$

where n is the interior unit normal to Γ.

Put

$$\left.\begin{aligned} \phi_0 &= \frac{1}{4}x_2^2 \\ \phi_1 &= 1 \\ \phi_2 &= 2x_1^2 - x_2^2 \\ \phi_3 &= 8x_1^4 - 24x_1^2 \cdot x_2^2 + 3x_2^4 . \end{aligned}\right\} \quad (3\text{-}50)$$

Note that these functions satisfy Eqs. 3-39 and 3-40.

The domain Ω and its boundary Γ are shown in Fig. 3-5. Partial derivatives with respect to the interior normal are shown.

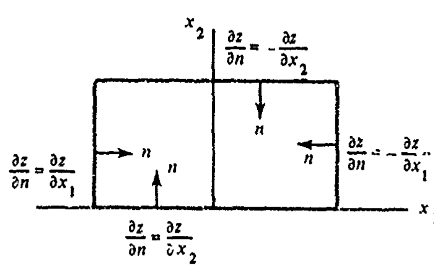


Figure 3-5. Boundary Condition for Example 3-4

The procedure is now to form

$$\bar{z} = \frac{1}{4}x_2^2 + c_1 + c_2(2x_1^2 - x_2^2) + c_3(8x_1^4 - 24x_1^2x_2^2 + 3x_2^4)$$

and, with the aid of the expressions for $\partial z/\partial n$ in Fig. 3-5, compute $\bar{z} + \partial\bar{z}/\partial n$ at L points around the boundary Γ. At a typical point, e.g., (1, 1/2),

$$\bar{z} + \frac{\partial \bar{z}}{\partial n} = \frac{1}{16} + c_1 - (11/4)c_2 - (285/16)c_3.$$

At this point it is required that

$$1/16 + (c_1^+ - c_1^-) - (11/4)(c_2^+ - c_2^-) - (285/16)(c_3^+ - c_3) < \gamma$$

and

$$-1/16 - (c_1^+ - c_1^-) + (11/4)(c_2^+ - c_2^-) - (285/16)(c_3^+ - c_3^-) < \gamma.$$

Similar inequalities in the c_i^+, c_i^-, and γ will be obtained at all other boundary points chosen. Under the requirements $c_i^+ > 0$, $c_i^- > 0$, and $\gamma > 0$, the problem of minimizing γ is then solved.

Rabinowitz in Ref. 1, page 141, reports that an approximate solution obtained by the above method is

$$c_1 = -0.5571,\quad c_2 = 0.0764,\quad c_3 = 0.0024,$$

$$\gamma = 0.0053.$$

This means that at all the boundary points x^{ℓ}, $|\,\bar{z} + \partial\bar{z}/\partial n\,| < 0.0053$. A result called a maximum principle from the theory of second-order elliptic partial differential equations then implies

$$|\,z(x) - \bar{z}(x)\,| < 0.0053,\quad x \text{ in } \Omega$$

where $z(x)$ is the true solution of Eqs. 3-48 and 3-49. This powerful result guarantees that the approximate solution $\bar{z}$ generated by linear programming is within 0.0053 of the true solution throughout Ω.

REFERENCES

1. P. Rabinowitz, "Applications of Linear Programming to Numerical Analysis", *SIAM Review*, Vol. 10, No 2, p. 121, April 1968.

2. H. Weyl, "The Elementary Theory of Convex Polyhedra", in H. W. Kuhn and A. W. Tucker, *Contributions to the Theory of Games*, Annals of Mathematical Studies No. 24, Princeton University Press, 1950.

3. G. Owen, *Game Theory*, Saunders, Philadelphia, 1968.

4. T. L. Saaty, *Mathematical Methods of Operations Research*, McGraw-Hill, New York, 1959

5. W. S. Dorn, R. E. Gomory, and H. J. Greenberg, "Automatic Design of Optimal Structures", *Journal de Mecanique*, Vol. 3, No. 1, March 1964, pp. 25-52.

6. W. R. Spillers and J. Farrell, "On the Analysis of Structural Design", *J Math. Anal. Appl*, Vol. 25, 1969, pp. 285–295.

CHAPTER 4

NONLINEAR PROGRAMMING AND FINITE DIMENSIONAL OPTIMAL DESIGN

4-1 INTRODUCTION TO THE THEORY OF NONLINEAR PROGRAMMING (NLP)

As pointed out in the preceding chapter, inequality constraints play a central role in engineering design problems. The inequalities treated in Chapter 3, however, are of a rather special form, namely, they involve only linear functions of the variables of the problem. It is a rare real-world design problem which can be put into this form. In general, the inequality constraints as well as the cost or return function in real-world problems are nonlinear. For this reason, a more general theory than that presented in Chapter 3 is needed.

The class of problems considered here is called nonlinear programming, or mathematical programming. A vast amount of literature has been devoted to this class of problems in recent years. Several books on the subject which contain reviews of this literature are Refs. 1, 2, and 3. In view of this extensive literature, the purpose of this paragraph is simply to state the nonlinear programming problem and present some key results needed in the study of methods of optimal design.

4-1.1 NONLINEAR PROGRAMMING PROBLEMS

For convenience and clarity in the development of methods of solution, the nonlinear programming problem will be stated in two forms. The first form is given by Definition 4-1.

Definition 4-1: The first nonlinear programming problem NLP, is: find $x \in R^n$ to

$$\left.\begin{aligned} &\text{minimize } f(x) && (4\text{-}1)\\ &\text{subject to} \\ &g(x) < 0 && (4\text{-}2)\end{aligned}\right\} \text{NLP}$$

$$\text{where} \quad g(x) = \begin{bmatrix} g_1(x) \\ \vdots \\ g_m(x) \end{bmatrix}.$$

Unless otherwise specified, it will be assumed that $f(x)$ and $g(x)$ are continuously differentiable. Other than this differentiability requirement, $f(x)$ and $g(x)$ are as general as required for a particular problem.

A second form of nonlinear programming problem, which may actually be included in NLP, is given by Definition 4-2.

Definition 4-2: The second nonlinear programming problem NLP′, is: find $x \in R^n$ to

$$\left.\begin{aligned} &\text{minimize } f(x) && (4\text{-}3)\\ &\text{subject to} \\ &g(x) < 0, && (4\text{-}4)\\ &\text{and} \\ &h(x) = 0 && (4\text{-}5)\end{aligned}\right\} \text{NLP}'$$

where $g(x) = \begin{bmatrix} g_1(x) \\ \vdots \\ g_m(x) \end{bmatrix}$,

and

$$h(x) = \begin{bmatrix} h_1(x) \\ \vdots \\ h_p(x) \end{bmatrix}.$$

Unless otherwise specified, it will be assumed that $f(x)$, $g(x)$, and $h(x)$ are continuously differentiable.

Very much as in the linear programming problem, the points x which satisfy the constraints of NLP and NLP′ are characterized by Definition 4-3.

Definition 4-3: The sets of points $x \in R^n$ that satisfy the constraints NLP and NLP′ are called constraint sets. They are denoted

$$D = \{x \in R^n \mid g(x) \leq 0\}$$

for NLP, and

$$D' = \{x \in R^n \mid g(x) \leq 0 \text{ and } h(x) = 0\}$$

for NLP′.

For convenience, Theorem 2-2, which was stated previously in Chapter 2, is given here (Theorem 4-1) as it applies to nonlinear programming problems.

Theorem 4-1: If $f(x)$ is continuous on D (D') and this set is closed and bounded in R^n, then NLP (NLP′) has a solution which is an absolute minimum of $f(x)$ in D (D').

This theorem is one of the most easily obtained yet most powerful results in optimization theory. It guarantees existence of a solution with only very mild assumptions. This result is a consequence of properties of R^n. In the infinite dimensional optimization problems of Chapter 6, the space of variables lacks these properties so that no analogous result is available.

Theorem 4-2 provides an easy test for closedness of the constraint set.

Theorem 4-2: If the functions $g(x)$ and $h(x)$ are continuous, then the sets D and D' are closed in R^n.

The boundedness hypothesis of Theorem 4-1 may be more difficult to check, particularly in complex problems. One must show that there exists a number a such that if $x \in D$ or D', then $x^T x < a$.

To see that NLP′ can actually be included in NLP, define

$$g_{i+m}(x) = h_i(x), i = 1, \ldots, p$$

and

$$g_{i+m+p}(x) = -h_i(x), i = 1, \ldots, p.$$

Now, NLP′ is equivalent to the NLP:

minimize $f(x)$

subject to

$\hat{g}(x) \leq 0$,

where $\hat{g}(x) = \begin{bmatrix} g_1(x) \\ \vdots \\ g_{m+2p}(x) \end{bmatrix}$.

This is true since

$$g_i(x) \leq 0, \ i = m+1, \ldots, m+2p$$

is just

$$h_j(x) < 0,\ j = 1, \ldots, p$$

and

$$-h_j(x) < 0,\ j = 1, \ldots, p$$

which is equivalent to

$$h(x) = 0.$$

It should be clear that problems of maximizing $\hat{f}(x)$ are put into the form NLP or NLP′ simply by defining $f(x) = -\hat{f}(x)$. Further, constraints of the form $\hat{g}(x) > 0$ are transformed to the proper form simply by defining $g(x) = -\hat{g}(x)$. These transformations involve no theoretical or practical difficulty. As will be seen in par. 4-2, even though the transformation of NLP′ into NLP involves no theoretical difficulty, severe practical difficulties occur. The explicit characterization of equality constraints in NLP′ will be useful later, when methods of constructing solutions are discussed.

Comparing nonlinear programming problems with the unconstrained problems of Chapter 2, one might conclude that the nature of the cost function $f(x)$ will determine the location of the minimum point, with only a check required to verify that constraints are satisfied. Since the linear programming problem is a special case of the nonlinear programming problem, the results of Chapter 3 show vividly that this conclusion is false. In the linear problem, the cost function plays only a minor role in the simplex algorithm and most of the computational effort is expended operating on the constraint functions.

While results from the linear programming problem yield valuable insight into the nonlinear programming problem, one must be careful not to generalize too much. To illustrate some differences between linear and nonlinear programming, two examples will now be treated.

Example 4-1:

Minimize

$$f(x) = (x_1 - 3)^2 + (x_2 - 3)^2$$

subject to constraints

$$-x_1 < 0$$

$$-x_2 < 0$$

$$x_1 + x_2 - 4 < 0.$$

The constraint set is the shaded triangular region in Fig. 4-1.

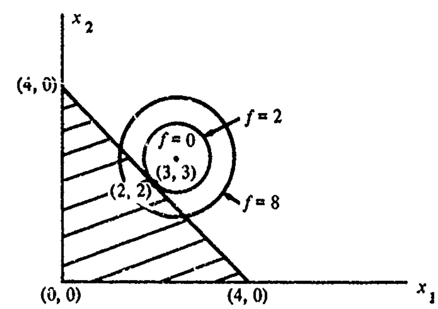


Figure 4-1. Graphical Solution of Example 4-1

If the constraints are ignored, $f(x)$ takes on its minimum at the point (3,3). Observing the circles, which are plots of constant value curves of $f(x)$, it is clear that the smallest value $f(x)$ takes on in the shaded triangle is $f(2,2) = 2$. This is, therefore, the solution of the problem.

It should be noted that even though the

solution occurred on the boundary of the constraint set, it did not occur at a corner as it would have if the problem had been linear.

Example 4-2:

Minimize

$$f(x) = (x_1 - 1)^2 + (x_2 - 1)^2$$

subject to the constraints

$$-x_1 \leq 0$$

$$-x_2 \leq 0$$

$$x_1 + x_2 - 4 \leq 0.$$

The constraint set is just the same as in the previous problem. The cost function, however, has been modified.

If the constraints are ignored, $f(x)$ takes on its minimum at (1,1). Since this combination of design variables satisfies the constraints, it is the solution of Example 4-2. The solution of this nonlinear programming problem, therefore, occurs in the interior of the constraint set. This behavior contrasts sharply with that of linear programming problems where the solution must occur on the boundary of the constraint set.

These examples show conclusively that the properties of NLP, and hence, also NLP′, differ considerably from those of LP.

Theoretical results and computational methods for NLP and NLP′ will also be more complex than those for the linear programming problem. The reason for this is clear. Strong use was made of linearity of the functions involved in the linear programming problem, and this linearity is not present in the nonlinear programming problem. The increased complexity of nonlinear as opposed to the linear problems is not surprising since increased complexity generally accompanies this transition in all mathematical disciplines.

Due to the complexity of NLP and NLP′, methods of obtaining their solutions are generally computational in nature. Moreover, in many meaningful engineering problems, convergence proofs are not available so the designer must depend heavily on his engineering intuition. One must be extremely careful in applying engineering intuition to certain aspects of optimization problems, however. In most problems of engineering analysis, existence and uniqueness of solutions are taken for granted since these properties hold for very general classes of problems such as linear elasticity, dynamics, circuit theory, and structural analysis. Existence and uniqueness questions in optimization problems are, however, by no means trivial. For instance, before the designer commits himself to a design based on an optimum obtained by a computational algorithm, he should seriously consider the possibility that this optimum is only relative and an absolute optimum exists that will give much better results.

Due to the weakness of intuition in dealing with optimization problems and the inherent complexity of these problems, the importance of theoretical results concerning existence, uniqueness, and necessary and sufficient conditions cannot be overemphasized. The remainder of this paragraph and par. 4-2 are devoted to these questions, while pars. 4-3 through 4-5 contain methods for obtaining solution of NLP and NLP′.

4-1.2 GLOBAL THEORY

In nonlinear programming problems one often obtains a relative minimum of $f(x)$ in the constraint set. The question arises, "Is this relative minimum an absolute minimum?" In general problems it is difficult to answer this question. There is a class of problems, however, in which this question is easily answered. This class is described by Definition 4-4.

Definition 4-4: If D (D') is a convex set and $f(x)$ is convex on D (D') then NLP (NLP′) is called a convex programming problem.

Theorem A-1, Appendix A, guarantees that if $g_i(x)$, $i = 1, ..., m$, are convex functions, then the set D is convex. Since the equalities (Eq. 4-5) in NLP′ define a surface in R^n, it is clear that D' is the intersection of that surface with the set $\{x \in R^n | g_i(x) \leq 0, i = 1, ..., m\}$. The surface is convex if and only if it is a plane, or equivalently, if and only if each $h_j(x)$ is linear in x. Since by Theorem A-6, Appendix A, the intersection of two convex sets is convex, D' is convex if $g_i(x)$, $i = 1, ..., m$, are convex and $h_j(x)$, $j = 1, ..., p$ are linear. The class of problems NLP′ which are convex is, therefore, quite restricted.

As will be clear from what follows, convexity is a very desirable property. However, in the real world, many optimization problems are nonconvex. In spite of this fact, the study of convex problems is justified. Many results which hold only in convex problems have led to constructive methods which are effective for finding local extrema in nonconvex problems. Some of these methods would probably never have been developed if only general nonconvex problems had been treated.

One of the powerful results which follows due to convexity is given in Theorem 4-3.

Theorem 4-3: A relative minimum in a convex programming problem is an absolute minimum.

4-1.3 LOCAL THEORY

Without convexity it is difficult to say much about global properties of the solution of NLP or NPL′. Considerable theory is available, however, which characterizes local minima. The approach in the local theory is to suppose that $f(x)$ has a relative minimum at a point in D or D' and then find conditions on $f(x)$, $g(x)$, and $h(x)$ which must hold at this point. In this way, many points in D and D' may be eliminated as candidates for a relative extrema and perhaps relative extrema can even be located using these conditions. Such conditions, therefore, are called "necessary". In some problems it will be possible to obtain a set of conditions that, if satisfied at a point, guarantee that this point yields a relative extremum. Conditions of this kind, of course, are called "sufficient".

As often happens in engineering, the engineer needs a powerful result developed in mathematics to solve his problem. Proof of this result, however, may be very complex and, in fact, contribute very little to the engineer's insight into his problems. This appears to be the case in many phases of optimization theory, in particular, in the study of necessary and sufficient conditions in nonlinear programming. In the remainder of this paragraph results will be borrowed from mathematical developments.

Before meaningful results may be given for NLP and NLP′, the following conditions will be required of the constraint functions $g(x)$ and $h(x)$.

Definition 4-5: (First-order constraint qualification): Let x^0 be a point in the constraint set D' (or D if there are no equality constraints) and let the functions $g(x)$ and $h(x)$ be differentiable at x^0. Then the first-order constraint qualification holds at x^0 if for any nonzero $y \in R^n$ such that $\nabla g_i(x^0)y < 0$ for each i with $g_i(x^0) = 0$ and $\nabla h(x^0)y = 0$, then y is tangent to a differentiable arc passing from x^0 into the constraint set.

Geometrically, this definition says that if the vector y is a direction which, to first order, appears to point from x^0 into the constraint set, then there is a curve with y as tangent which actually passes from x^0 into the constraint set. The conditions $\nabla g_i(x^0)y < 0$ for $g_i(x^0) = 0$ and $\nabla h(x^0) = 0$ are just first order perturbations of $g_i(x)$ and $h(x)$ which indicate that a small move in the y-direction ought to do the right thing to $g_i(x)$ and $h(x)$. This is illustrated in Fig. 4-2.

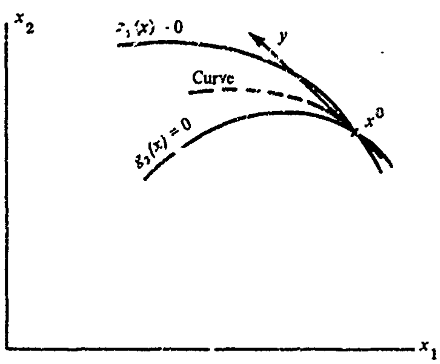


Figure 4-2. First-order Constraint Qualification

While all constraints do not satisfy the first-order constraint qualification, the following theorem (Ref. 1, page 19) identifies a class of constraints which do.

Theorem 4-4: If $g(x)$ and $h(x)$ are differentiable at x^0 in D' and if the gradients $\nabla g_i(x^0)$, for i with $g_i(x^0) = 0$ and $\nabla h_j(x^0)$ are linearly independent, $j = 1, \ldots, p$, then the first-order constraint qualification is satisfied.

In this result, and in fact, in the remainder of this paragraph, the problem NLP′ is described. It is clear, however, that putting $p = 0$ in NLP′ yields NLP. One of the principal results of nonlinear programming may now be stated. For proof the reader is referred to Ref. 1, page 20.

Theorem 4-5: (Kuhn-Tucker Necessity Theorem): Let the functions $f(x)$, $g(x)$, and $h(x)$ be differentiable and let the constraint functions satisfy the first-order constraint qualifications at a point $\bar{x}$ in D' of NLP′. In order that $\bar{x}$ be a relative minimum for NLP′ it is necessary that there exist multipliers $v \in R^m$ and $w \in R^p$ such that

$$v_i \geq 0, i = 1, \ldots, m \tag{4-6}$$

$$v_i g_i(\bar{x}) = 0, i = 1, \ldots, m \tag{4-7}$$

and

$$\nabla L(\bar{x}, v, w) = 0 \tag{4-8}$$

where

$$L(x, v, w) = f(x) + v^T g(x) + w^T h(x) \tag{4-9}$$

is called the Lagrangian.

In a sense, Theorem 4-5 is an existence theorem. It asserts that if $\bar{x}$ yields a relative minimum for NLP′, then the multipliers v and w exist and that Eq. 4-8 is satisfied. Occasionally, one will run across an argument attempting to justify this theorem which states that

$$f(x) = f(x) + v^T g(x) + w^T h(x) = L(x, v, w)$$

since v is defined by Eq. 4-7 and $h(x) = 0$. It is then claimed that since $\bar{x}$ yields a relative minimum for $f(x)$ it must yield a relative minimum for $L(x,\nu,w)$, so $\nabla L(x,\nu,w) = 0$ must hold. This argument is *not* valid. For a rigorous proof of Theorem 4-5 the reader is referred to Ref. 1.

Theorem 4-6 states additional conditions which are required to hold if the functions appearing in NLP' have two derivatives.

Theorem 4-6: (Second-order Necessary Conditions): Let $f(x)$, $g(x)$, and $h(x)$ have two continuous derivatives at a point $\bar{x}$ in D'. Further, let the vectors $\nabla g_i(\bar{x})$, for all i with $g_i(\bar{x}) = 0$, and $\nabla h(\bar{x})$ be linearly independent. If $\bar{x}$ yields a relative minimum for NLP', then it is necessary that there exist ν and w satisfying Eqs. 4-6, 4-7, and 4-8. Further, for every $y \epsilon R^n$ such that $\nabla g_i(\bar{x})y = 0$ when $g_i(\bar{x}) = 0$, and $\nabla h(\bar{x})y = 0$, it is necessary that

$$y^T \nabla^2 L(\bar{x},\nu,w)y \geq 0 \tag{4-10}$$

For proof of this theorem, see Ref. 1, page 25. Note that the existence of ν and w satisfying Eqs. 4-6, 4-7, and 4-9 is a consequence of Theorem 4-5. Even though this theorem involves second-order conditions, it still gives only necessary conditions.

A theorem which gives conditions which, if satisfied at some point, are sufficient to guarantee that this point yields a relative minimum for NLP' will now be stated. For proof of this theorem, see Ref. 1, page 30.

Theorem 4-7: (Second-order Sufficient Conditions): Let $f(x)$, $g(x)$, and $h(x)$ be twice differentiable functions at a point $\bar{x}$. If for $x \epsilon D'$ there exist ν and w satisfying

$$\nu_i \geq 0, \; i = 1, \ldots, m$$

$$\nu_i g_i(\bar{x}) = 0, \; i = 1, \ldots, m$$

$$\nabla L(\bar{x},\nu,w) = 0$$

and if for every nonzero $y \epsilon R^n$ such that $\nabla g_i(\bar{x})y = 0$ for $\nu_i > 0$, $\nabla g_i(\bar{x})y \leq 0$ for $g_i(\bar{x}) = 0$ and $\nu_i = 0$, and $\nabla h(\bar{x})y = 0$, it is true that

$$y^T \nabla^2 L(\bar{x},\nu, w)y > 0 \tag{4-11}$$

then $\bar{x}$ yields an isolated relative minimum for NLP'.

It should be noted that there is a gap between the sufficient conditions of Theorem 4-7 and the necessary conditions of Theorem 4-6. Strict inequality is required in Eq. 4-11 for a larger set of vectors y that may yield only equality in Eq. 4-10. It is doubtful that a single, tractable set of conditions exist that are both necessary and sufficient for the general problem NLP'.

There is one class of nonlinear programming problems in which conditions may be given that are both necessary and sufficient for an absolute extremum. This class is the convex programming problem.

Theorem 4-8: Let $f(x)$ and $g_i(x)$, $i = 1, \ldots, m$, be continuously differentiable and convex, then necessary and sufficient conditions for $\bar{x}$ to be an absolute minimum point of NLP are that there exists $\nu \epsilon R^n$ such that

$$g(\bar{x}) \leq 0$$

$$\nu_i g_i(\bar{x}) = 0, \; i = 1, \ldots, m$$

$$\nu_i \geq 0, \; i = 1, \ldots, m$$

and

$$\nabla f(\bar{x}) + \sum_{i=1}^{m} \nu_i \nabla g_i(\bar{x}) = 0.$$

The technical presentation of par. 4-1 ends with this satisfying result. Several comments are, however, appropriate at this point. The analytic necessary and sufficient conditions of par. 4-1 could be used to construct solutions of NLP by solving systems of nonlinear equations. This is particularly true of the results of Theorem 4-8. If one reads the current literature, however, he is led to the distinct conclusion that iterative methods based on successive improvements are too effective to bypass in favor of methods that require solution of complicated, nonlinear, algebraic equations.

Even if the results of par. 4-1 are never used by the designer to construct solutions of nonlinear programming problems, they are still very powerful tools. Verification of the hypotheses of one of the theorems may mean the difference between going onto the computer with the comforting knowledge that a unique solution exists as opposed to the frustrating experience of having computer print-out which may be meaningless.

4-2 THEORY OF FINITE DIMENSIONAL OPTIMAL DESIGN

The nonlinear programming problems of par. 4-1 are quite general and may be applied to a variety of optimization problems. As is frequently the case with very general formulations of problems, special features of some problems within the class being studied are not exploited. This appears to be the case when general nonlinear programming theory is applied to solve optimal design problems. Interpretation of certain of the variables and constraints in the problem NLP′, in the context of optimal design, yields very effective computational methods of solution. This paragraph will be devoted to stating the finite dimensional optimal design problem, drawing an analogy with NLP′, and stating necessary and sufficient conditions that follow directly from the theorems stated in the preceding paragraph.

4-2.1 FINITE DIMENSIONAL OPTIMAL DESIGN PROBLEMS

The class of problems to be treated in this paragraph is, in a sense, a special case of the nonlinear programming problem NLP′. However, by developing a theory for the new class of problems which takes advantage of its special features, a more efficient solution algorithm may be obtained.

The general optimal design problem must have several of the features of NLP′. Namely, it is required to have a cost (return) function which is to be minimized (maximized) and a set of constraints that describe the performance demanded of the object being designed. It is in the representation of constraints that the optimal design problem differs from NLP′.

In most problems of design in the real-world the object being designed is required to behave according to some law of physics. This behavior is described analytically by a set of variables called state variables. Further, there is a second set of variables that describe the object itself rather than its behavior. These variables are called design variables since they are to be chosen by the designer so that the object being designed performs its required function. It generally happens that the laws of physics that determine the state variables depend on the design variables so the two sets of variables are related.

To illustrate the difference between state and design variables, consider the following design problems.

1. Find the coefficient of damping in an

automobile shock absorber so that peak acceleration in the passenger compartment due to road conditions is as small as possible.

The coefficient of damping is the design variable since it describes the object being designed, and its magnitude is to be fixed by the designer. Acceleration on the other hand is a state variable since it describes the behavior of the object being designed. Further, this state variable may be determined by Newton's laws of motion. Note that the designer has no direct control over the state variable. He may effect it only indirectly by adjusting the design variable. This is typical of state and design variables.

2. Determine the size of beams to be used in a structure so that when a given set of loads are applied stresses are within certain given limits, the deflection of certain points on the structure is within given limits, and the structure is as light in weight as possible.

Beam sizes are the design variables in this problem since they describe the structure being designed and they must be chosen by the designer. Stress and deflection, however, are state variables that are determined by equilibrium and force deflection relations. Again, the designer has no direct control over stress and deflection. He may effect these quantities only by varying the size of beams in the structure.

In most real-world design problems the state and design variables are clearly identified. In what follows, the state variable will be an n-vector, $z \in R^n$, and the design variable will be a k-vector, $b \in R^k$. The basic elements of the optimal design problem are described by Definition 4-6

Definition 4-6 The finite dimensional optimal design problem (OD) is a problem of determining $b \in R^k$ to

$$\text{minimize } f(z,b) \tag{4-12}$$

subject to (OD)

$$h(z,b) = 0 \tag{4-13}$$

$$\phi(z,b) \leq 0 \tag{4-14}$$

where

$$h(z,b) = \begin{bmatrix} h_1(z,b) \\ \vdots \\ h_n(z,b) \end{bmatrix}, \quad \phi(z,b) = \begin{bmatrix} \phi_1(z,b) \\ \vdots \\ \phi_m(z,b) \end{bmatrix}, \tag{4-15}$$

and all the functions of the problem are required to have first-order derivatives. Further, it is required that the $(n + k)$ vectors

$$\left[\frac{\partial \phi_i}{\partial z}, \frac{\partial \phi_i}{\partial b}\right] \tag{4-16}$$

are linearly independent for all i with $\phi_i(z,b) = 0$ and that the matrix

$$\frac{\partial h}{\partial z} \tag{4-17}$$

is nonsingular.

The assumption that the matrix $\frac{\partial h}{\partial z}$ is nonsingular guarantees, by the implicit function theorem (Ref 4, page 181), that for given b there is a unique solution of Eq 4-13 for z. Further the state variable determined from Eq. 4-13 as a function of b, is differentiable with respect to b. This fact will be needed later when constructive methods are developed.

4-2.3 LOCAL THEORY

Since it is very seldom that the state equations (Eq. 4-13) are linear in both z and b, convexity of the constraint set and hence the problem will be rare. For this reason, no global results based on convexity will be discussed. In case Eq. 4-13 is linear, however, global results may be obtained by applying the Theorems 4-3 and 4-8.

It is clear that if a new variable $x \in R^{n+k}$ is defined as

$$x = \begin{bmatrix} z \\ b \end{bmatrix} \tag{4-18}$$

then the problem OD may be put into the form NLP'. According to Theorem 4-4, the first-order constraint qualification will be satisfied for OD (with $x \in R^{n+k}$ as independent variable) if the row vectors

$$\left[\frac{\partial h_i}{\partial z}, \frac{\partial h_i}{\partial b}\right], i = 1, ..., n \tag{4-19}$$

$$\left[\frac{\partial \phi_j}{\partial z}(z,b), \frac{\partial \phi_j}{\partial b}(z,b)\right],$$

$$\text{for } j \text{ with } \phi_j(z,b) = 0 \tag{4-20}$$

are linearly independent. Theorem 4-5 may now be applied to the problem OD.

Theorem 4-9 (First-order Necessary Conditions) Let all the functions appearing in OD be differentiable at a point $\bar{z}, \bar{b}$ which satisfies Eqs. 4-13, 4-14, and 4-15. Further let the vectors (Eqs. 4-19, 4-20, and 4-21) be linearly independent at $\bar{z},\bar{b}$. Then there exist multipliers $\lambda \in R^n$ and $\mu \in R^m$, with $\mu \geq 0$ such that for

$$H = f(z,b) + \lambda^T h(z,b) + \mu^T \phi(z,b) \tag{4-21}$$

$$\frac{\partial H}{\partial b}(\bar{z},\bar{b}) = 0 \tag{4-22}$$

$$\frac{\partial H}{\partial z}(\bar{z},\bar{b}) = 0 \tag{4-23}$$

and

$$\mu_j \phi_j(\bar{z},\bar{b}) = 0, \; j = 1, ..., m. \tag{4-24}$$

The proof of this theorem may be constructed by simply writing down the necessary conditions of Theorem 4-5 in terms of x and then separating the components of x as in Eq. 4-18.

In exactly the same way the second-order necessary and sufficient conditions of Theorems 4-6 and 4-7, respectively, may be stated for the problem OD. No essential simplification of the statements of those theorems occurs, however, so the theorems are not restated here.

Theorem 4-9, just as Theorem 4-5, is difficult to use in constructing solutions of OD. Considerable difficulty arises because one does not know which of the inequalities in OD is an equality. For problems with a small number of inequality constraints this may not be a difficult obstacle, particularly if the designer has a good intuitive idea of which constraints will be equalities. If, on the other hand, there are a large number of inequality constraints, then the number of combinations of constraints which may be equalities is large. It is, therefore, difficult to determine just which combinations will be equalities. An analytic solution is extremely difficult in this case.

Rather than attempt to use the necessary conditions to construct candidate solutions, a more direct approach will be followed. The remainder of this chapter will be devoted to direct methods of solving NLP, NLP′, and OD.

4-3 SEQUENTIALLY UNCONSTRAINED MINIMIZATION TECHNIQUES (SUMT)

A favorite method of solving difficult problems, particularly among mathematicians, is to reduce a difficult problem to a sequence of easy problems. Each of the easy problems is solved and if the method is any good, the sequence of solutions of easy problems will converge to the solution of the difficult problem. As the title might imply, SUMT follows just this pattern. It should be clear that a central part of this method must be results which guarantee convergence, at least in cases where solutions are known to exist.

The method presented here essentially reduces NLP and NLP′ to a sequence of auxiliary problems which may be solved by the methods of Chapter 2. The cost function of NLP or NLP′ is augmented by a function called a penalty function. The penalty function is formed from the constraint functions in such a way that as a parameter approaches zero (or perhaps infinity) the unconstrained minimum of the augmented cost function converges to the solution of NLP or NLP′. Two basically different ways of doing this are presented here. Each has its computational and theoretical advantages and disadvantages that will be described later.

Due to the large body of theory concerning SUMT, results will be presented in this paragraph without proof. The reader is referred for proofs and an extended discussion of SUMT to the complete and well-written text of Fiacco and McCormick (Ref. 1). Theoretical results guaranteeing convergence are presented here to indicate the level of the known theory of SUMT, rather than as a complete treatment of the subject.

4-3.1 INTERIOR METHOD

The interior SUMT is based on the idea of using the constraint functions to erect a barrier at the boundary of the constraint set D of NLP by adding a penalty function to $f(x)$ which approaches infinity as the boundary of D is approached from the interior. Once the solution of the augmented problem is obtained, the penalty function is altered so as to effect $f(x)$ less in the interior of D. This behavior is illustrated in Fig. 4-3.

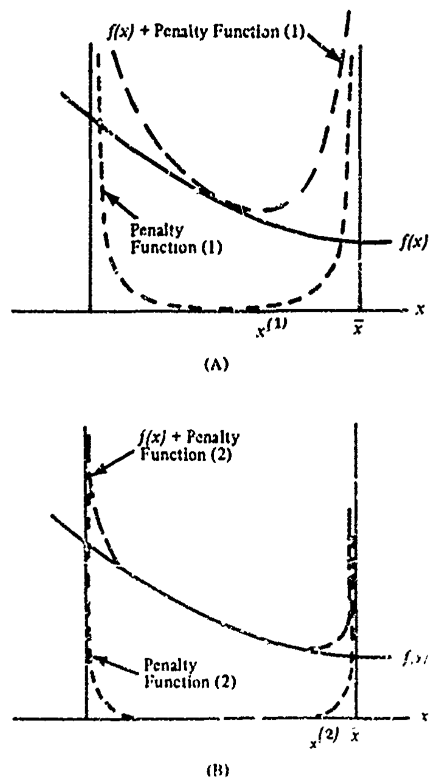


Figure 4-3. Penalty Functions

As illustrated in Fig. 4-3, when the penalty function is decreased on the interior of D, the minimum of the second augmented cost function $x^{(2)}$ is closer to the solution $\bar{x}$ than the minimum of the first augmented cost function $x^{(1)}$. The idea, of course, is that the sequence of points $x^{(i)}$ generated in this way converges to $\bar{x}$.

It should be clear why this approach is discussed only for NLP and not NLP′. The constraint set of NLP′ can have no interior due to the equality constraints. It is possible that NLP has no interior and in this case the interior SUMT is not applicable. In what follows, it is assumed that the constraint set D of NLP has an interior.

The sequence of points $x^{(i)}$ which is to converge to the minimum point is generated by minimizing

$$f(x) + S(r_i)\, I(x) \tag{4-25}$$

without regard to constraints, where $S(r_i)\, I(x)$ is continuous for x in the interior of D and $S(r_i)\, I(\hat{x}) = +\infty$ for any $\hat{x}$ such that $g_j(\hat{x}) = 0$ for any $1 \leq j \leq m$. It is clear that if one begins an iterative minimization technique of Chapter 2 at a point in the interior of D, then a relative minimum point will be found which must lie in the interior of D. Otherwise, the minimizing sequence would have had to climb over a portion of the auxiliary cost surface that is infinitely high and none of the methods will do this.

In order to obtain the sequence of points $x^{(i)}$, the parameter r_i is allowed to approach zero. To insure that the sequence $x^{(i)}$ converges to a relative minimum point, the functions $I(x)$ and $S(r)$ are required to have the following properties:

1. $I(x)$ is continuous and non-negative on the interior of the constraint set D and if $\{x^k\}$ is any sequence of points in E^n converging to $\bar{x}$ where $g_j(\bar{x}) = 0$ for some j, then $\lim_{k\to\infty} I(x^k) = +\infty$.

2. $S(r)$ is continuous and if $r_1 > r_2 > 0$, then $S(r_1) > S(r_2)$ 0 and if r_i is a sequence of numbers converging to zero, then $\lim_{i\to\infty} S(r_i) = 0$.

Probably the most common penalty functions $I(x)$ and $S(r)$ are

$$I(x) = -\sum_{j=1}^{m} \frac{1}{g_j(x)} \tag{4-26}$$

and

$$S(r) = r. \tag{4-27}$$

Any pair of functions satisfying properties No. 1 and No. 2 associated with Eq. 4-25, however, is suitable. It may be to the designer's advantage to choose another form for any particular problem. For other suitable choices of penalty functions, see Ref. 1, page 68.

The algorithm for solving NLP by the interior point technique is given in Definition 4-7.

Definition 4-7. The interior point sequentially unconstrained minimization algorithm is given by the following:

Step 1. Define the function

$$U(x,r) = f(x) + S(r)\, I(x). \tag{4-28}$$

where $S(r)$ and $I(x)$ satisfy properties No. 1 and No. 2. Choose $r_0 > 0$ and $x^{(0)}$ in the interior of the constraint set D.

Step 2. Beginning at $x^{(0)}$ minimize $U(x,r_0)$ without regard to constraints to obtain $x^{(1)}$. Any of the methods of Chapter 2 may be employed for this purpose.

Step 3. For $l = 0, 1, 2, \ldots$, choose $r_{l+1} > 0$ such that $r_{l+1} < r_l$. Beginning at $x^{(l)}$ minimize $U(x, r_{l+1})$ without regard to constraints to obtain $x^{(l+1)}$, where l is the iteration index.

Step 4. As $r_l \to \infty$, if $\| x^{(l+1)} - x^{(l)} \|$ and $| f[x^{(l+1)}] - f[x^{(l)}] |$ are sufficiently small, terminate the process and take $x^{(l+1)}$ as the solution of NLP. Otherwise return to Step 3.

In order to be sure that this algorithm will lead to a solution of NLP, one would like to have a result that as $r_k \to 0$, a solution is approached. Such a result is contained in Theorem 4-10.

Theorem 4-10: In the interior point algorithm just given let:

$f(x)$, $g_1(x), \ldots, g_m(x)$ be continuous on the constraint set D, (4-29)

$S(r)$ and $I(x)$ satisfy properties No. 1 and No. 2, (4-30)

The interior of D be nonempty, (4-31)

There be a relative minimum point $\bar{x}$ in D such that $f(\bar{x}) < f(x)$ for all $x \neq \bar{x}$ in some neighborhood of $\bar{x}$, where $\bar{x}$ is not an isolated point of D, (4-32)

$\{ r_l \}$ be a strictly decreasing sequence which converges to zero. (4-33)

Then for $x^{(0)}$ sufficiently near $\bar{x}$ and r_0 sufficiently small,

$$\lim_{l \to \infty} x^{(l)} = \bar{x}. \tag{4-34}$$

Further,

$$\lim_{l \to \infty} S(r_l)\, I[x^{(l)}] = 0 \tag{4-35}$$

$$\lim_{l \to \infty} f[x^{(l)}] = \lim_{l \to \infty} U[x^{(l)}, r_l] = f(\bar{x}) \tag{4-36}$$

$\{ f[x^{(l)}] \}$ is monotone decreasing (4-37)

and

$\{ I[x^{(l)}] \}$ is monotone increasing. (4-38)

For proof of this theorem see Ref. 1, page 47.

It has been noted throughout the previous development that if NLP is convex – i.e., $f(x)$, $g_1(x), \ldots, g_m(x)$ are convex – then "nice" things happen. One of these "nice" things is given in Theorem 4-11.

Theorem 4-11: If NLP is convex with a unique minimum point $\bar{x}$, $g_j(x)$, $j = 1, \ldots, m$, are twice continuously differentiable, and if Eqs 4-29 through 4-33 hold, then $x^{(l)}$ generated by the given algorithm will converge to the minimum point.

It should be noted that Step 1 of the algorithm (Def. 4-7) required a point $x^{(0)}$ in the interior of the constraint set but no method of obtaining such a point was given. This question will be addressed later in this paragraph.

Example 4-3: Solve the LP

$$f(x_1, x_2) = x_1 + x_2 = \text{minimum}$$

$$g_1(x_1, x_2) = -x_1 < 0$$

$$g_2(x_1, x_2) = -x_2 < 0$$

using the interior point SUMT.

Solution:

$$U(x,r) = x_1 + x_2 - r\left[-\frac{1}{x_1} - \frac{1}{x_2}\right].$$

The functions $f(x)$, $g_1(x)$, and $g_2(x)$ are convex and by Theorem A-5, Appendix A, so are $-1/g_1(x)$ and $1/g_2(x)$. Since $r > 0$, $U(x,r)$ is convex and thus has a unique minimum. To find it, put

$$\frac{\partial U}{\partial x_1} = 0 = 1 - \frac{r}{(x_1)^2}$$

$$\frac{\partial U}{\partial x_2} = 0 = 1 - \frac{r}{(x_2)^2}$$

so

$$x_1 = r^{1/2}$$

$$x_2 = r^{1/2}$$

As $r \to 0$, $x_1 \to 0$ and $x_2 \to 0$ so the solution of Example 4-3 is

$$(x_1, x_2) = (0, 0)$$

4-3.2 EXTERIOR METHOD

Unlike the interior method, starting points for the exterior SUMT are not required to be in the constraint set of NLP. The basic idea in the exterior method is to add to the cost function a penalty function that is positive for points outside the constraint set and zero inside the constraint set. This, in effect, discourages the minimum of the new augmented cost function from being too far from the constraint set if the original cost function $f(x)$ is "well behaved" outside the constraint set. It is clear that this approach may not be taken if $f(x)$ is undefined or takes on negatively infinite values outside the constraint set. One very appealing aspect of the exterior method is that it handles equality as well as inequality constraints without difficulty, so that it can be used on NLP'.

The penalty function employed for the exterior method will have the form

$$P(t)\ E(x) \tag{4-39}$$

where $P(t)$ and $E(x)$ are required to satisfy the conditions:

1. $E(x) = 0$ if x is in the constraint set, and $E(x) > 0$ if x is outside the constraint set.

2. $P(t)$ is continuous and if $t_2 > t_1 > 0$, then $E(t_2) > E(t_1) > 0$. Further, if $t_i \to +\infty$ then $\lim_{i\to\infty} P(t_i) = +\infty$.

Probably the most common choice for $P(t)$ and $E(x)$ is

$$P(t) = t \tag{4-40}$$

and

$$E_1(x) = \sum_{j=1}^{m} [g_j(x) + |g_j(x)|]^2 \tag{4-41}$$

and

$$E_2 = \sum_{j=1}^{P} [h_j(x)]^2$$

where

$$E = E_1 + E_2.$$

The basic idea for the exterior method was given by R. Courant in 1943 (Ref. 5). He argued that if

$$T(x,t) = f(x) + P(t)\,E(x) \qquad (4\text{-}42)$$

were minimized without regard to constraints using t_1 and t_2 with $t_2 > t_1$, then since the augmented cost function is penalized more when t_2 is used than when t_1 is used, the minimum point corresponding to t_2 should be closer to the constraint set and hence, closer to the minimum point of $f(x)$ on the constraint set.

An explicit algorithm for solving NLP or NLP′ by this method is given in Definition 4-8.

Definition 4-8: The exterior point sequentially unconstrained minimization algorithm is given by the following:

Step 1. Make an engineering estimate $x^{(0)}$ of the solution of NLP or NLP′.

Step 2. Choose $t_1 > 0$ and beginning at $x^{(0)}$ find an unconstrained minimum point of

$$T(x,t_1) = f(x) + P(t_1)\,E(x)$$

denoted $x^{(1)}$.

Step 3. Continue with $i = 2, \ldots$ by choosing $t_i > t_{i-1}$ and starting from $x^{(i-1)}$ obtain an unconstrained minimum point of

$$T(x,t_i) = f(x) + P(t_i)\,E(x)$$

denoted $x^{(i)}$.

Step 4. As $t_i \to \infty$, if $\| [x^{(i)} - x^{(i-1)}] \|$ and $| f[x^{(i)}] - f[x^{(i-1)}] |$ are sufficiently small, terminate the process and take $x^{(i)}$ as the solution of NLP. Otherwise, return to Step 3.

Very much as in the case of the interior method, Theorem 4-12 guarantees a certain measure of success.

Theorem 4-12: In the exterior point algorithm let:

$f(x)$, $g_1(x)$, ..., $g_m(x)$ be continuous for all x. (4-43)

$E(x)$ and $P(t)$ satisfy conditions No. 1 and No. 2 of Eq. 4-39. (4-44)

There be a relative minimum point $\bar{x}$ in that admissible domain D such that $f(\bar{x}) < f(x)$ for all $x \neq \bar{x}$ in some neighborhood of $\bar{x}$, where $\bar{x}$ is not an isolated point of D. (4-45)

The sequence $\{t_i\}$ is strictly increasing to $+\infty$. (4-46)

Then for $x^{(0)}$ sufficiently close to $\bar{x}$, and t_i sufficiently large,

$$\lim_{i\to\infty} x^{(i)} = \bar{x} \qquad (4\text{-}47)$$

$$\lim_{i\to\infty} P(t_i)E[x^{(i)}] = 0 \qquad (4\text{-}48)$$

$$\lim_{i \to \infty} f[x^{(i)}] = f(\bar{x}) \tag{4-49}$$

$$\lim_{i \to \infty} T[x^{(i)}, t_i] = f(\bar{x}) \tag{4-50}$$

$$\{ f[x^{(i)}] \} \text{ is monotone decreasing} \tag{4-51}$$

$$\{ E[x^{(i)}] \} \text{ is monotone decreasing.} \tag{4-52}$$

For proof of this theorem, see Ref. 1, page 57.

Very much as in the interior method, if the NLP or NLP′ is convex, then convergence is guaranteed by Theorem 4-13.

Theorem 4-13: If NLP or NLP′ is convex with a unique minimum point, and if Eqs. 4-43, 4-44, and 4-46 hold, then regardless of the estimates $x^{(0)}$ and t_i, the sequence $x^{(i)}$ generated by the algorithm given by Theorem 4-12 will converge to the minimum point.

Example 4-4: Solve

$$f(x_1, x_2) = x_1^2 + 2x_2^2 = \text{minimum}$$

$$h(x_1, x_2) = x_1 + x_2^2 - 1 = 0$$

by the exterior point SUMT.

$$T(x,t) = x_1^2 + 2x_2^2 + t(x_1 + x_2 - 1)^2$$

$$\frac{\partial T}{\partial x_1} = 2x_1 + 2t(x_1 + x_2 - 1) = 0$$

$$\frac{\partial T}{\partial x_2} = 4x_2 + 2t(x_1 + x_2 - 1) = 0$$

Subtracting,

$$4x_2 - 2x_1 = 0, \text{ or } x_1 = 2x_2.$$

Then

$$4x_2 + 2t(3x_2 - 1) = 0$$

and

$$x_2 = \frac{t}{2 + 3t}.$$

As

$$t \to \infty,\ x_2 \to \frac{1}{3} \text{ and } x_1 \to \frac{2}{3}.$$

The solution is then

$$(x_1, x_2) = \left(\frac{1}{3}, \frac{2}{3} \right).$$

4-3.3 MIXED INTERIOR-EXTERIOR METHOD

Both the interior and the exterior methods presented in pars. 4-3.1 and 4-3.2 are not applicable in certain kinds of problems. In particular, the interior method cannot be used if the interior of the constraint set is empty, such as in the case with equality constraints. The exterior method cannot be used if some constraint function is not defined or is ill-behaved outside the constraint. A combination of the two methods will now be given which allows the treatment of problems which may have both these undesirable features and thus could not be treated by either pure interior or exterior methods.

For convenience, consider NLP′

$$\text{minimize } f(x) \tag{4-53}$$

subject to

$$g_i(x) \le 0,\ i = 1, \ldots, m \tag{4-54}$$

$$h_j(x) = 0,\ j = 1, \ldots, p \tag{4-55}$$

where the set of all points which satisfy the m inequalities Eq. 4-54, has an interior. As might be expected, the constraints, Eq. 4-54, will be dealt with using an interior point penalty function and the constraints, Eq. 4-55, will be dealt with using an exterior point penalty function.

The penalty function used here will be

$$S(r_i)\, I(x) + P(t_i)\, E(x),$$

where $S(r)I(x)$, $P(t)$, and $E(x)$ satisfy conditions No. 1 and No. 2 preceding Eq. 4-26 and No. 1 and No. 2 of Eq. 4-39. It is understood that $I(x)$ is a function of only the constraint functions in Eq. 4-54 and $E(x)$ is a function of only those in Eq. 4-55. A general minimizing algorithm for NLP is now given in Definition 4-9.

Definition 4-9: The mixed interior-exterior sequentially unconstrained minimization algorithm is given by the following:

Step 1. Make an engineering estimate $x^{(0)}$ of the solution of NLP′.

Step 2. Choose $r_1 > 0$ and $t_1 > 0$ and obtain an unconstrained minimum of

$$V(x,r_1,t_1) = f(x) + S(r_1)\, I(x) + P(t_1)\, E(x), \quad (4\text{-}56)$$

denoted $x^{(1)}$.

Step 3. Continue with $i = 2, \ldots$ by choosing $r_i < r_{i-1}$ and $t_i > t_{i-1}$ and starting from $x^{(i-1)}$ finding an unconstrained minimum point of

$$V(x,r_i,t_i) = f(x) + S(r_i)\, I(x) + P(t_i)\, E(x) \quad (4\text{-}57)$$

denoted $x^{(i)}$.

Step 4. As $r_i \to 0$ and $t_i \to +\infty$, if $\|x^{(i)} - x^{(i-1)}\|$ and $|f[x^{(i)}] - f[x^{(i-1)}]|$ are sufficiently small, terminate the process and take $x^{(i)}$ as the solution of NLP′. Otherwise return to Step 3.

As might be expected from a study of the two methods which were combined to form the mixed method, a convergence result is given by Theorem 4-14.

Theorem 4-14: In the mixed point algorithm let:

$g_1(x), \ldots, g_m(x)$ be continuous on the nonempty interior of their constraint set and $f(x), h_1(x), \ldots, h_p(x)$ be continuous for all x. (4-58)

$S(r)$, $I(x)$, $P(t)$, and $E(x)$ satisfy conditions No. 1 and No. 2 preceding Eq. 4-26, and No. 1 and No. 2 of Eq. 4-39. (4-59)

There exist a relative minimum point $\bar{x}$ in the admissible domain D' of Eqs. 4-54 and 4-55 combined, such that $f(\bar{x}) < f(x)$ for all $x \neq \bar{x}$ in some neighborhood of $\bar{x}$, where $\bar{x}$ is not an isolated point of D'. (4-60)

The sequence $\{r_i\}$ be strictly decreasing to 0 and $\{t_i\}$ be strictly increasing to $+\infty$. (4-61)

Then for $x^{(0)}$ sufficiently close to $\bar{x}$, r_i sufficiently small, and t_i sufficiently large,

$$\lim_{i\to\infty} S(r_i)\, I[x^{(i)}] = 0 \quad (4\text{-}62)$$

$$\lim_{i \to \infty} P(t_i) E[x^{(i)}] = 0 \tag{4-63}$$

and

$$\lim_{i \to \infty} V[x^{(i)}, r_i, t_i] = f(\bar{x}). \tag{4-64}$$

For proof, see Ref. 1, page 60.

Example 4-5: Solve

$f(x_1, x_2) = -x_1 + x_2 =$ minimum

$g(x_1, x_2) = -\ell n\, x_2 < 0$

$h(x_1, x_2) = x_1 + x_2 - 1 = 0.$

Solution:

Since $g(x_1, x_2)$ is unbounded at $x_2 = 0$, it must be treated by the interior method and since $h(x_1, x_2) = 0$ prevents the constraint set from having an interior, it must be treated by the exterior method. From

$$V(x,r,t) = -x_1 + x_2 + \frac{r}{\ell n x_2} + t(x_1 + x_2 - 1)^2 \tag{4-65}$$

$$\frac{\partial V}{\partial x_1} = -1 + 2t(x_1 + x_2 - 1) = 0 \tag{4-66}$$

$$\frac{\partial V}{\partial x_2} = 1 - \frac{r}{x_2 \ell n^2 x_2} + 2t(x_1 + x_2 - 1) = 0 \tag{4-67}$$

Subtracting Eq. 4-67 from Eq. 4-66,

$$\frac{r}{x_2 \ell n^2 x_2} = 2$$

so

$$x_2 = e^{\left[\frac{r}{2x_2}\right]^{1/2}}$$

Since $g(x_1, x_2) < 0$ is satisfied at all times, $x_2 > 1$. Taking the limit as $r \to 0$, then, $x_2 \to 1$.

As $t \to +\infty$, it is necessary that $x_1 + x_2 - 1 \to 0$ or Eq. 4-66 will be violated. Therefore, in the limit $x_1 = 0$.

The minimum point is, therefore, $(x_1, x_2) = (0,1)$.

4-3.4 DETERMINATION OF AN INTERIOR POINT

In order to begin the interior point or the mixed interior-exterior point algorithm, it is necessary to have a point $x^{(0)}$ which satisfies a certain set of inequalities, i.e., a point interior to a given constraint set. Let this set of inequalities be

$$g_i(x) < 0, i = 1, \ldots, m. \tag{4-68}$$

If there are other inequalities or equalities which will be treated by the exterior point method, they are ignored for now.

Let $y^{(0)}$ be a first estimate of an interior point of the set defined by Eq. 4-68. Denote the inequalities of Eq. 4-68 which are not strictly satisfied by

$$N = \{i \mid g_i[y^{(0)}] \geq 0\}$$

and the inequalities which are strictly satisfied by

$$K = \{i \mid g_i[y^{(0)}] < 0\}.$$

The object now is to move from $y^{(0)}$ to points where successively more inequalities move from N into K; i.e., so that no constraint which was previously satisfied is violated, but constraints which were previously violated are satisfied. This may be accomplished by minimizing the unconstrained cost function

$$U(y,r_j) = \sum_{i \in N} g_i(y) + S(r_j) \sum_{i \in K} I[g_i(y)]$$

with respect to y, where $S(r)$ and $I(x)$ satisfy the four constraints – No. 1 and No. 2 preceding Eq. 4-26, and No. 1 and No. 2 of Eq. 4-39 – and r_j is a strictly decreasing sequence. The result is denoted $y^{(j)}$.

As soon as $y^{(j)}$ is such that $g_i[y^{(j)}] < 0$ for some i previously in N, that constraint function is switched to K. In this way, constraint functions from N may get to K but those in K may never fall back to N. Once all the constraints in N are switched to K, the process is stopped and the resulting $y^{(j)}$ is in the interior of the constraint set of Eq. 4-68. If the minimum of $U(y,r_j)$ is found as $r_j \to 0$ and there are still constraints in N, then the constraint set defined by Eq. 4-68 has no interior. In this case, NLP is infeasible (has no solution) or certain of the constraints of Eq. 4-68 will have to be treated by exterior point methods.

4-4 STEEPEST DESCENT METHODS FOR NLP

In Chapter 2 a gradient method is presented for finding the minimum of an unconstrained function. Such a direct method has properties that make it attractive and worth developing for the solution of NLP. It is clear, however, that due to constraints the gradient method studied earlier does not apply directly to NLP. It is the object of this paragraph to develop a method which uses only first derivative information to make successive improvements in an estimated solution of NLP. A study of the problem NLP′ will be better included in the next paragraph.

Geometrically, the method presented here will first investigate the direction of most rapid decrease in the cost function $f(x)$. As seen in par. 2-4, this is $-\nabla f^T(x)$. This direction is then projected onto the tangent hyperplane to the boundary of the constraint set at x. A small move in the resulting direction will then decrease $f(x)$ and will not cause excessive violation of constraints. This process is repeated as long as $f(x)$ may be decreased.

Instead of basing the derivation of the method on a geometric argument, the work will all be done analytically. The reason for this is twofold. First, geometric ideas in higher dimensions are not always as clear as those in two and three dimensions. Second, the analytical method used here will be employed in deriving algorithms in continuous problems where geometric concepts are much more difficult.

Extensive use will be made of matrix calculus notation in this paragraph.

Recall that for $g(x) = \begin{bmatrix} g_1(x) \\ \vdots \\ g_m(x) \end{bmatrix}$, $x \in R^n$

$$\frac{\partial g}{\partial x} = \left[\frac{\partial g_i}{\partial x_j}\right]_{m \times n}$$

Further, the symbol

$$\delta x = \begin{bmatrix} \delta x_1 \\ \vdots \\ \delta x_n \end{bmatrix}$$

will denote a change in x and a δ in front of a quantity which depends on x will denote the first order change in that quantity due to the change δx in x. For example, for the scalar function $f(x)$,

$$\delta f(x) = \frac{\partial f}{\partial x}\,\delta x.$$

Note that this first-order change is just the first term in a Taylor expansion, Ref. 4, page 84, of $f(x)$, so $\delta f(x)$ is an accurate approximation of the change in $f(x)$ only for small δx.

The method to be developed here resembles an interior method in the sense of par. 4-3. Therefore, the method of generating an interior point (one which satisfies all the constraints) presented in par. 4-3 may be utilized to obtain a starting point. It is assumed now that his has been done, and that an estimate $x^{(0)}$ of the solution of NLP is available which satisfies

$$g(x^{(0)}) < 0.$$

4-4.1 THE DIRECTION OF STEEPEST DESCENT

If the point $x^{(0)}$ is in the interior of the constraint set, then the gradient method of par. 2-4 applies and the direction in which $x^{(0)}$ should be altered is

$$\delta x = -k\,\frac{\partial f^T}{\partial x}\,[x^{(0)}]\,, \qquad (4\text{-}69)$$

$k > 0$.

In the remaining case, the point $x^{(0)}$ is on the constraint boundary so $g_l[x^{(0)}] = 0$ for some l. For convenience define the set

$$A(x) = \{\, l \mid g_l(x) = 0 \,\}\,, \qquad (4\text{-}70)$$

i.e., the collection of indices of constraint functions which are equalities at the point x.

The object is now to find a direction of change $\delta\bar{x}$, $\delta\bar{x}^T\delta\bar{x} = 1$, such that $\delta x = k\delta\bar{x}$ for sufficiently small $k > 0$ will decrease $f(x)$ without violating any constraints. The problem is then to find $\delta\bar{x}$ such that

$$\delta f = \frac{\partial f}{\partial x}\,[x^{(0)}]\,\delta\bar{x}$$

is minimum subject to

$$\delta g_l[x^{(0)}] = \frac{\partial g_l}{\partial x}\,[x^{(0)}]\,\delta\bar{x} < 0,$$

$$l \in A[x^{(0)}]$$

and

$$\delta\bar{x}^T\delta\bar{x} = 1.$$

For further convenience, define the column vector of constraint functions which are zero as

$$\bar{g}(x) = \begin{bmatrix} g_l(x) \\ l \in A[x^{(0)}] \end{bmatrix}.$$

In this notation the problem is

$$\text{minimize } \frac{\partial f}{\partial x}[x^{(0)}]\,\delta\bar{x} \qquad (4\text{-}71)$$

subject to

$$\frac{\partial \bar{g}}{\partial x}[x^{(0)}]\,\delta\bar{x} < 0, \qquad (4\text{-}72)$$

$$\delta\bar{x}^T\delta\bar{x} = 1. \qquad (4\text{-}73)$$

It is assumed that at points where several $g_l(x) = 0$, the gradients are linearly independent. This is enough to satisfy the first-order constraint qualification (Theorem 4-4) for the constraint, (Eq. 4-72). Theorem 4-5, therefore, may be applied to obtain the necessary conditions

$$\frac{\partial f}{\partial x} + \tilde{\lambda}^T \frac{\partial \tilde{g}}{\partial x} + 2\lambda_0 \delta \tilde{x}^T = 0, \qquad (4\text{-}74)$$

where the components, $\tilde{\lambda}_i > 0$, of $\tilde{\lambda}$ correspond to g_i with indices in $A[x^{(0)}]$.

Assume for the time being that $\delta g_i = 0$ for all $i \in A[x^{(0)}]$. Then taking the transpose of Eq. 4-74 and premultiplying by $\partial \tilde{g}/\partial x$ yields

$$\frac{\partial \tilde{g}}{\partial x}\frac{\partial f^T}{\partial x} + \frac{\partial \tilde{g}}{\partial x}\frac{\partial \tilde{g}^T}{\partial x}\tilde{\lambda} + 2\lambda_0 \frac{\partial \tilde{g}}{\partial x}\delta \tilde{x} = 0,$$

or since $\delta \tilde{g} = 0$,

$$\frac{\partial \tilde{g}}{\partial x}\frac{\partial f^T}{\partial x} + \frac{\partial \tilde{g}}{\partial x}\frac{\partial \tilde{g}^T}{\partial x}\tilde{\lambda} = 0.$$

Since the gradients of $g_i[x^{(0)}]$ for $i \in A[x^{(0)}]$ are assumed linearly independent, the coefficient matrix of $\tilde{\lambda}$ is nonsingular and

$$\tilde{\lambda} = -\left(\frac{\partial \tilde{g}}{\partial x}\frac{\partial \tilde{g}^T}{\partial x}\right)^{-1}\frac{\partial \tilde{g}}{\partial x}\frac{\partial f^T}{\partial x}. \qquad (4\text{-}75)$$

If all components of $\tilde{\lambda}$ are non-negative, then the assumption that all $\delta g_i = 0$, $i \in A[x^{(0)}]$ is valid and $\delta \tilde{x}$ which solves the problems of Eqs. 4-71, 4-72, and 4-73, is obtained directly from Eqs. 4-74 and 4-75. On the other hand, if $\tilde{\lambda}_i < 0$ for some $i \in A[x^{(0)}]$, then this component of g is removed from $\tilde{g}$. Equivalently, $A[x^{(0)}]$ is redefined as

$$\hat{A}[x^{(0)}] = \{\, i \mid g_i[x^{(0)}] = 0 \text{ and } \tilde{\lambda}_i > 0 \,\}$$

and $\tilde{g}[x^{(0)}]$ is redefined as

$$\hat{g}[x^{(0)}] = \begin{bmatrix} g_i[x^{(0)}] \\ i \in \hat{A} \end{bmatrix}$$

With this new $\hat{g}$, Eq. 4-74 yields

$$\delta \hat{x} = -\frac{1}{2\lambda_0}\left[I - \frac{\partial \hat{g}^T}{\partial x}\left(\frac{\partial \hat{g}}{\partial x}\frac{\partial \hat{g}^T}{\partial x}\right)^{-1}\frac{\partial \hat{g}}{\partial x}\right]_{[x^{(0)}]} \frac{\partial f^T}{\partial x}[x^{(0)}]. \qquad (4\text{-}76)$$

Note that $\lambda_0 > 0$ is required since if $A[x^{(0)}]$ is empty, then Eq. 4-76 must reduce to the negative gradient direction.

Putting

$$P = \left(I - \frac{\partial \hat{g}^T}{\partial x}[x^{(0)}] \times \left\{\frac{\partial \hat{g}}{\partial x}[x^{(0)}]\frac{\partial \hat{g}^T}{\partial x}[x^{(0)}]\right\}^{-1} \times \frac{\partial \hat{g}}{\partial x}[x^{(0)}]\right) \qquad (4\text{-}77)$$

Eq. 4-76 becomes

$$\delta \hat{x} = -\frac{1}{2\lambda_0} P \frac{\partial f^T}{\partial x}[x^{(0)}].$$

Substituting this into Eq. 4-73,

$$\frac{1}{(2\lambda_0)^2}\frac{\partial f}{\partial x}[x^{(0)}] P^T P \frac{\partial f^T}{\partial x}[x^{(0)}] = 1.$$

Solving for $1/2\lambda_0$, $\delta \hat{x}$ becomes

$$\delta \hat{x} = -\left\{\frac{\partial f}{\partial x}[x^{(0)}] P^T P \frac{\partial f^T}{\partial x}[x^{(0)}]\right\}^{-1/2} \times P \frac{\partial f^T}{\partial x}[x^{(0)}]. \qquad (4\text{-}78)$$

Eq. 4-78 gives a unit vector $\delta\hat{x}$ in the constrained direction of steepest descent at $x = x^{(0)}$. The problem is now one of determining just how big a step should be taken, i.e.,

$$\delta x = k\delta\hat{x} \tag{4-79}$$

where $k > 0$ must be chosen.

Before the problem of step size is treated, however, an algorithm may be stated for determination of the direction of steepest descent, namely:

Step 1. Using the method of par. 4-3, obtain an estimate of the solution of NLP, $x^{(0)}$, which is in the constraint set.

Step 2. Let $l > 0$ denote the number of the present iteration. Compute $g_i[x^{(l)}]$, $j = 1, \ldots, m$, and form the set $A[x^{(l)}]$. Compute $\partial f/\partial x[x^{(l)}]$ and $\partial \bar{g}_i/\partial x[x^{(l)}]$ for $i \in A[x^{(l)}]$.

Step 3. Compute $\tilde{\lambda}$ in Eq. 4-75. For all $\tilde{\lambda}_i < 0$, delete i from $A[x^{(l)}]$ to form $\hat{A}[x^{(l)}]$.

Step 4. Compute P in Eq. 4-77 and δx in Eq. 4-78. If $P = 0$, then this is the solution of NLP.

Example 4-6: Compute the direction of steepest descent at the point (2,2) for the NLP

$$\text{minimize } f(x_1, x_2) = (x_2)^2 - x_1$$

$$g_1(x_1, x_2) = x_1 - x_2 < 0$$

$$g_2(x_1, x_2) = -x_1 - x_2 + 2 < 0.$$

First,

$$g_1(2, 2) = 0$$

$$g_2(2, 2) = -2.$$

Therefore, $A[x^{(1)}] = \{1\}$. As required by Step 2 of the Algorithm

$$\frac{\partial f}{\partial x}(2, 2) = [-1, 4]$$

$$\frac{\partial g_1}{\partial x}(2, 2) = [1, -1]$$

By Step 3,

$$\tilde{\lambda} = -\left([1, -1]\begin{bmatrix}1\\-1\end{bmatrix}\right)^{-1} \times [1, -1]\begin{bmatrix}-1\\4\end{bmatrix} = \frac{5}{2} > 0$$

so $\hat{A} = A$. For Step 4

$$P = \begin{bmatrix}1 & 0\\0 & 1\end{bmatrix} - \begin{bmatrix}1\\-1\end{bmatrix}\left([1, -1]\begin{bmatrix}1\\-1\end{bmatrix}\right)^{-1} \times [1, -1] = \begin{bmatrix}1/2 & 1/2\\1/2 & 1/2\end{bmatrix}$$

Finally,

$$\delta\hat{x} = \left([1, -4]\begin{bmatrix}1/2 & 1/2\\1/2 & 1/2\end{bmatrix}\begin{bmatrix}1/2 & 1/2\\1/2 & 1/2\end{bmatrix}\begin{bmatrix}-1\\4\end{bmatrix}\right)^{-1/2} \times \begin{bmatrix}1/2 & 1/2\\1/2 & 1/2\end{bmatrix}\begin{bmatrix}-1\\4\end{bmatrix} = -\frac{\sqrt{2}}{3}\begin{bmatrix}3/2\\3/2\end{bmatrix}.$$

4-4.2 STEP SIZE DETERMINATION

There are many techniques presented in the literature for determining the size of step to be taken in the constrained direction of steepest descent. Three of these techniques are presented here. The first technique applies to a specialized class of problems in which the constraint functions are linear. The second and third methods apply to the general nonlinear problem.

4-4.2.1 ROSEN'S METHOD FOR LINEAR CONSTRAINTS

If the constraint functions are linear, then once the direction of steepest descent $\delta\hat{x}$ is found, it may be followed without leaving the constraint boundary until a constraint $g_i(x) = 0$, for i not in $\hat{A}[x^{(j)}]$. This algorithm, therefore, can lead to rather long step sizes.

Constraints here are restricted to the form

$$G_i^T x - b_i < 0, \; i = 1, ..., m$$

where G_i is an $n \times 1$ matrix of constants. The step size is to be determined so that k is as small as possible and still

$$G_i^T [x^{(j)} + k\delta\hat{x}] - b_i = 0$$

for some $i \notin A[x^{(j)}]$. Only those i need to be considered for which $G_i^T\delta\hat{x} > 0$, since otherwise this constraint can never go from strict inequality to equality. The step size k, therefore, is chosen as

$$k = \min_{\substack{G_i^T\delta\hat{x} > 0 \\ i \notin \hat{A}[x^{(j)}]}} \left[\frac{b_i - G_i^T x^{(j)}}{G_i^T\delta\hat{x}}\right]$$

The point $x^{(j+1)}$ therefore is given by

$$x^{(j+1)} = x^{(j)} + k\delta\hat{x}$$

if $\partial f/\partial x[x^{(j+1)}]\delta\hat{x} \leqslant 0$. The process in the algorithm for the direction of steepest descent is repeated from Step 2, and a new step size k is computed as above. If, on the other hand, $\partial f/\partial x[x^{(j+1)}]\delta\hat{x} > 0$, then a relative minimum has been bypassed. To locate this relative minimum, do a one-dimensional search in the direction $\delta\hat{x}$ starting at $x^{(j)}$ to obtain $x^{(j+1)}$.

This process may be summarized in *Rosen's Algorithm:*

Step 1. Compute

$$k = \min_{\{G_i^T\delta\hat{x} = 0\}} \left[\frac{b_i - G_i^T x^{(j)}}{G_i^T\delta\hat{x}}\right].$$

Step 2. Compute

$$\frac{\partial f}{\partial x}[x^{(j)} + k\delta\hat{x}]$$

If

$$\frac{\partial f}{\partial x}[x^{(j)} + k\delta\hat{x}]\delta\hat{x} < 0$$

then

$$x^{(j+1)} = x^{(j)} + k\delta\hat{x}$$

and go to Step 4.

Step 3. If

$$\frac{\partial f}{\partial x}[x^{(j)} + k\delta\hat{x}]\delta\hat{x} > 0,$$

then find $\bar{k}$ so as to minimize

$$f[x^{(j)} + k\delta\hat{x}]$$

4-4.2 STEP SIZE DETERMINATION

There are many techniques presented in the literature for determining the size of step to be taken in the constrained direction of steepest descent. Three of these techniques are presented here. The first technique applies to a specialized class of problems in which the constraint functions are linear. The second and third methods apply to the general nonlinear problem.

4-4.2.1 ROSEN'S METHOD FOR LINEAR CONSTRAINTS

If the constraint functions are linear, then once the direction of steepest descent $\delta\hat{x}$ is found, it may be followed without leaving the constraint boundary until a constraint $g_i(x) = 0$, for i not in $\hat{A}[x^{(j)}]$. This algorithm, therefore, can lead to rather long step sizes.

Constraints here are restricted to the form

$$G_i^T x - b_i < 0, \; i = 1, \ldots, m$$

where G_i is an $n \times 1$ matrix of constants. The step size is to be determined so that k is as small as possible and still

$$G_i^T [x^{(j)} + k\delta\hat{x}] - b_i = 0$$

for some $i \notin A[x^{(j)}]$. Only those i need to be considered for which $G_i^T \delta\hat{x} > 0$, since otherwise this constraint can never go from strict inequality to equality. The step size k, therefore, is chosen as

$$k = \min_{\substack{\{G_i^T \delta\hat{x} > 0\} \\ \{i \notin \hat{A}[x^{(j)}]\}}} \left[\frac{b_i - G_i^T x^{(j)}}{G_i^T \delta\hat{x}} \right]$$

The point $x^{(j+1)}$ therefore is given by

$$x^{(j+1)} = x^{(j)} + k\delta\hat{x}$$

if $\partial f/\partial x[x^{(j+1)}]\delta\hat{x} \leq 0$. The process in the algorithm for the direction of steepest descent is repeated from Step 2, and a new step size k is computed as above. If, on the other hand, $\partial f/\partial x[x^{(j+1)}]\delta\hat{x} > 0$, then a relative minimum has been bypassed. To locate this relative minimum, do a one-dimensional search in the direction $\delta\hat{x}$ starting at $x^{(j)}$ to obtain $x^{(j+1)}$.

This process may be summarized in *Rosen's Algorithm:*

Step 1. Compute

$$k = \min_{\{G_i^T \delta\hat{x} = 0\}} \left[\frac{b_i - G_i^T x^{(j)}}{G_i^T \delta\hat{x}} \right].$$

Step 2. Compute

$$\frac{\partial f}{\partial x}[x^{(j)} + k\delta\hat{x}]$$

If

$$\frac{\partial f}{\partial x}[x^{(j)} + k\delta\hat{x}]\delta\hat{x} < 0$$

then

$$x^{(j+1)} = x^{(j)} + k\delta\hat{x}$$

and go to Step 4.

Step 3. If

$$\frac{\partial f}{\partial x}[x^{(j)} + k\delta\hat{x}]\delta\hat{x} > 0,$$

then find $\bar{k}$ so as to minimize

$$f[x^{(j)} + k\delta\hat{x}]$$

Put $x^{(j+1)} = x^{(j)} + \bar{k}\delta\hat{x}$ and go to Step 4.

Step 4. If $|f[x^{(j+1)}] - f[x^{(j)}]|$ and $\|x^{(j+1)} - x^{(j)}\|$ are sufficiently small, terminate and take $x^{(j+1)}$ as the solution of NLP. Otherwise, return to Step 2 of the constrained steepest-descent algorithm.

4-4.2.2 FIXED STEP WITH VARIABLE WEIGHTING

When the auxiliary problem, Eqs. 4-71 through 4-73, was formulated, it would have been possible to ask for the step size directly rather than just the direction of steepest descent. In many cases, the behavior of the solution is much more sensitive to changes in one variable than another. For stability of calculation then, rather than asking for a direction $\delta\bar{x}$ satisfying Eq. 4-73, the designer might request a change δx in $x^{(j)}$ which satisfies

$$\delta x^T W \delta x = \ell^2 \tag{4-80}$$

where W is a positive definite matrix (usually diagonal) and ℓ is a predetermined constant. The elements of W are often chosen so that expected changes in various components of x, δx, will contribute approximately the same magnitude to $\delta x^T W \delta x$. The matrix W, therefore, is chosen based on the designer's experience.

The analysis performed in obtaining the direction of steepest descent follows with only minor changes. The only changes of interest, computationally, are

$$\bar{\lambda} = -\left(\frac{\partial \bar{g}}{\partial x} W^{-1} \frac{\partial \bar{g}^T}{\partial x}\right)^{-1} \frac{\partial \bar{g}}{\partial x} W^{-1} \frac{\partial f^T}{\partial x} \tag{4-81}$$

$$P = W^{-1} \times \left[I - \frac{\partial \bar{g}^T}{\partial x}\left(\frac{\partial \bar{g}}{\partial x} W^{-1} \frac{\partial \bar{g}^T}{\partial x}\right)^{-1} \frac{\partial \bar{g}}{\partial x} W^{-1}\right]. \tag{4-82}$$

and

$$\delta x = -\left[\frac{\ell^2}{\left(\dfrac{\partial f}{\partial x} P^T W P \dfrac{\partial f^T}{\partial x}\right)}\right]^{1/2} P \frac{\partial f^T}{\partial x}. \tag{4-83}$$

Note that if the step size is made large, then considerable progress may be made toward the minimum point. However, since the constraint functions are nonlinear, violations may occur at any iteration. After a new point $x^{(j+1)}$ has been computed, the constraint functions should be checked. If any constraints are violated in excess of fixed tolerance, the method of par. 4-3 may be used to move $x^{(j+1)}$ back into the constraint set.

The computational method is then described in *Algorithm for Steepest Descent With Fixed Step Size:*

Step 1. Using the method of par. 4-3, obtain an interior estimate of the solution of NLP, $x^{(0)}$, which is in the constraint set. Further, choose the weighting matrix W and step size ℓ in Eq. 4-80.

Step 2. Let j denote the number of the present iteration. Compute $g_i[x^{(j)}]$ and form the set $A[x^{(j)}]$. Compute $\partial f/\partial x[x^{(j)}]$ and $\partial \bar{g}/\partial x[x^{(j)}]$ for $i \in A[x^{(j)}]$.

Step 3. Compute $\bar{\lambda}$ in Eq. 4-81. For all $\bar{\lambda}_i$

≤ 0, delete l from $A[x^{(j)}]$ to form $\hat{A}[x^{(j)}]$.

Step 4. Compute P in Eq. 4-82 and δx in Eq. 4-83.

Step 5. Compute $x^{(j+1)} = x^{(j)} + \delta x$. If any constraints are violated excessively, use the method of par. 4-3 to get from $x^{(j+1)}$ back into the constraint set.

Step 6. If $|f[x^{(j+1)}] - f[x^{(j)}]|$ and $\|x^{(j+1)} - x^{(j)}\|$ are sufficiently small, terminate. Otherwise, return to Step 2 (possibly with altered ℓ and W).

4-4.2.3 STEEPEST DESCENT WITH CONSTRAINT TOLERANCES

In par. 4-4.2.2 it was noted that a step may be made so large as to violate a constraint in excess of an admissible error. The method of choosing step size presented here will prevent this difficulty.

Let reasonable tolerances ϵ_l be assigned to constraint functions $g_l(x)$. The object here is to move in the constrained direction of steepest descent until some constraint function $g_l(x)$ violates the tolerance $g_l(x) > \epsilon_l$, or until a minimum of

$$f[x^{(j)} + k\delta x]$$

is reached.

A uniform step size in k may be chosen and steps taken, checking

$$g_l[x^{(j)} + k\delta x]$$

for each $l \in \hat{A}[x^{(j)}]$ and each step in k. The multiplier k is increased monotonically provided $f[x^{(j)} + k\delta x]$ is decreasing, and constraints do not exceed the given error tolerances. When either fails to hold, the resulting point is called $x^{(j+1)}$.

If the process is stopped because a constraint is violated in excess of its given tolerance, the method of par. 4-3 is used to obtain a new point in the constraint set and the process is repeated until the minimum point is located.

This method should be most effective when constraint functions are easily evaluated but derivatives are costly in computer time. The basic idea of the method is to prevent an excessive number of calculations of the constrained direction of steepest descent.

4-4.3 A STEEPEST DESCENT METHOD WITH CONSTRAINT ERROR COMPENSATION

In previous subparagraphs, steepest descent methods were given which at boundary points generated steps parallel to a constraint boundary in a direction which decreased the cost function as rapidly as possible. Due to nonlinearity of the constraint functions, and the finite step size, however, some constraints will invariably be violated. It is the object in this paragraph to present a new method motivated by the article (Ref. 6) which automatically corrects for violation in constraints.

Let $A[x^{(j)}] = \{ i | g_i[x^{(j)}] \geq 0 \}$ be the indices of constraint functions which are zero or are violated. As in the preceding development of this paragraph first-order Taylor approximations will be used to approximate functions appearing in NLP. The linearized version of NLP at an approximation to the solution, $x^{(j)}$, is

$$\text{minimize } \delta f = \frac{\partial f}{\partial x}[x^{(j)}]\,\delta x \tag{4-84}$$

subject to

$$\delta \bar{g} = \frac{\partial \bar{g}}{\partial x}[x^{(j)}]\,\delta x < \Delta \bar{g}, \tag{4-85}$$

where $\bar{g} = \begin{bmatrix} g_i[x^{(j)}] \\ i \in A \end{bmatrix}$ and $\Delta \bar{g}$ is taken as the desired change in $\bar{g}$, i.e., the total change taken at the designer's discretion. Usually, so long as the constraints are not violated excessively, the full violation may be corrected; i.e.

$$\Delta \bar{g}_i = -\bar{g}_i[x^{(j)}], \quad i \in A[x^{(j)}]. \tag{4-86}$$

In order that step size is not excessive, it is required that

$$\delta x^T \delta x = \ell^2 \tag{4-87}$$

where ℓ is small. Assuming Eq. 4-85 is an equality, necessary conditions for the linearized problem are obtained by using Theorem 4-5. From

$$L = \frac{\partial f}{\partial x}\,\delta x + \lambda^T \left(\frac{\partial \bar{g}}{\partial x}\,\delta x - \Delta \bar{g} \right)$$

$$+ \beta \delta x^T \delta x$$

and Theorem 4-5, it is necessary that $\lambda_i > 0$, and

$$\frac{\partial f^T}{\partial x} + \frac{\partial \bar{g}^T}{\partial x}\lambda + 2\beta\delta x = 0 \tag{4-88}$$

and

$$\lambda_i \left(\frac{\partial g_i}{\partial x}\,\delta x - \Delta g_i \right) = 0, \quad i \in A.$$

This set of equations is nonlinear in λ and x. Assuming Eq. 4-85 is an equality, then the necessary conditions reduce to only Eq. 4-88 and Eq. 4-85 as an equality. This system is linear and can be solved. The multiplier λ can then be determined and a check made to see whether all components are non-negative. If any component is negative, say k, then the assumption that Eq. 4-85 is an equality is violated and it may be concluded that the kth component of Eq. 4-85 should have been allowed to be a strict inequality. The index k is then deleted from A.

Premultiplying Eq. 4-88 by $\partial \bar{g}/\partial x$ and using Eq. 4-85 yields

$$\frac{\partial \bar{g}}{\partial x}\frac{\partial f^T}{\partial x} + \frac{\partial \bar{g}}{\partial x}\frac{\partial \bar{g}^T}{\partial x}\lambda + 2\beta\delta\bar{g} = 0.$$

It is assumed, as usual, that the gradients of all constraint functions which are zero or violated are linearly independent. Therefore, the coefficient matrix of λ is nonsingular and

$$\lambda = -\left(\frac{\partial \bar{g}}{\partial x}\frac{\partial \bar{g}^T}{\partial x} \right)^{-1} \times \left[\frac{\partial \bar{g}}{\partial x}\frac{\partial f^T}{\partial x} + 2\beta\Delta\bar{g} \right]. \tag{4-89}$$

Substituting Eq. 4-89 into Eq. 4-88 yields

$$\delta x = -\frac{1}{2\beta}\left[I - \frac{\partial \bar{g}^T}{\partial x}\left(\frac{\partial \bar{g}}{\partial x}\frac{\partial \bar{g}^T}{\partial x} \right)^{-1} \frac{\partial \bar{g}}{\partial x} \right] \frac{\partial f^T}{\partial x} + \frac{\partial \bar{g}^T}{\partial x}\left(\frac{\partial \bar{g}}{\partial x}\frac{\partial \bar{g}^T}{\partial x} \right)^{-1} \Delta \bar{g}. \tag{4-90}$$

This expression for δx could now be substituted into Eq. 4-87 to find β. To be more general, however, put $1/(2\beta) = \gamma > 0$ and define

$$\delta x^1 = -\left[I - \frac{\partial \bar{g}^T}{\partial x}\left(\frac{\partial \bar{g}}{\partial x}\frac{\partial \bar{g}^T}{\partial x}\right)^{-1}\frac{\partial \bar{g}}{\partial x}\right]\frac{\partial f^T}{\partial x} \qquad (4\text{-}91)$$

and

$$\delta x^2 = \frac{\partial \bar{g}^T}{\partial x}\left(\frac{\partial \bar{g}}{\partial x}\frac{\partial \bar{g}^T}{\partial x}\right)^{-1} \Delta \bar{g}\,. \qquad (4\text{-}92)$$

Using this new notation Eq. 4-90 becomes

$$\delta x = \gamma \delta x^1 + \delta x^2. \qquad (4\text{-}93)$$

This representation of δx has important properties given by Theorem 4-15.

Theorem 4-15: δx^1 and δx^2 of Eqs. 4-91 and 4-92 satisfy the conditions

1. $\delta x^{1^T} \delta x^2 = 0$

2. $\frac{\partial \bar{g}}{\partial x} \delta x^2 = \Delta \bar{g}$

3. $\frac{\partial \bar{g}}{\partial x} \delta x^1 = 0$

4. $\frac{\partial f}{\partial x} \delta x^1 < 0$

A method of choosing γ still has not been given. This parameter is interpreted as a step-size and may be determined by one-dimensional search or any other scheme chosen by the designer. In different applications, different methods have proved effective. No single scheme has been found that seems best. The choice of γ at this time constitutes an art as much as a science.

The use of this method may now be summarized in the *Steepest-descent Algorithm With Constraint Error Compensation.*

Step 1. Make an engineering estimate of the solution of NLP.

Step 2. Let the iteration number be $j > 0$. Compute $g_i[x^{(j)}]$ and form $A[x^{(j)}]$ and $\bar{g}$.

Step 3. Compute $\partial \bar{g}/\partial x[x^{(j)}]$ and $\partial f/\partial x[x^{(j)}]$ and choose the desired change $\Delta \bar{g}$ in g.

Step 4. Compute δx^1 and δx^2 in Eqs. 4-91 and 4-92.

Step 5. Choose γ by a suitable scheme. Calculate λ in Eq. 4-89. If any components λ_i are less than zero for $g_i[x^{(j)}]$ which are close to zero, remove these components from g and return to Step 3. If all $\lambda_i > 0$, proceed.

Step 6. Form

$$\delta x = \gamma \delta x^1 + \delta x^2$$

and

$$x^{(j+1)} = x^{(j)} + \delta x.$$

Step 7. If $|f[x^{(j+1)}] - f[x^{(j)}]|$ and $\|\delta x\|$ are sufficiently small, terminate the process. Otherwise return to Step 2.

4-5 STEEPEST DESCENT SOLUTION OF THE FINITE DIMENSIONAL OPTIMAL DESIGN PROBLEM

In this paragraph a steepest-descent method of solution of the problem OD is developed.

In many ways, the method of this paragraph is similar to the method of par. 4-4. Here, however, a distinction is made between design and state variables, and the two types of variables are treated quite differently.

The problem to be solved here is, just as in par. 4-2: Choose $b \in R^k$ and $z \in R^n$ to minimize

$$\left.\begin{aligned} &f(z, b) && (4\text{-}94)\\ &\text{subject to}\\ &h(z, b) = 0 && (4\text{-}95)\\ &\text{and}\\ &\phi(z, b) < 0 && (4\text{-}96)\end{aligned}\right\} \text{OD}$$

where $h(z,b) = [h_1(z,b), ..., h_n(z,b)]^T$, and $\phi(z,b) = [\phi_1(z,b), ..., \phi_m(z,b)]^T$. The state equations, Eq. 4-95, are put into vector form here in order to take advantage of the compact matrix calculus notation.

The steepest descent algorithm for OD is developed here by first approximating the nonlinear elements of OD by linear expressions in the various variables. The difference between the method presented here and that of par. 4-4 lies in the treatment of the state variable. In a sense, the state variable is a nuisance since it does not really describe the system being designed. The algorithm presented here is obtained by first eliminating the state variable from the linearized problem and then solving an explicit problem for an optimum improvement in the design variable.

Very much as in par. 4-4, an engineering estimate of the optimum design is made. It is denoted by $b^{(0)}$. Then the state equations, Eq. 4-95, are solved for the corresponding state $z^{(0)}$. Any method of analysis may be used to solve Eq. 4-95 for z. The object now is to determine a change in $b^{(0)}$, denoted δb, such that

$$b^{(1)} = b^{(0)} + \delta b \tag{4-97}$$

will be an "improved" design. The meaning of "improved" will be made clear as the analysis progresses. If the new design variable $b^{(1)}$ were substituted into Eq. 4-95, this equation could be solved for the corresponding new state variable $z^{(1)}$. Since the matrix $\partial h/\partial z[z^{(0)}, b^{(0)}]$ is nonsingular, the implicit function theorem, Ref. 4, page 181, guarantees that if $\| \delta b \|$ is small, then $z^{(1)} - z^{(0)}$ will be small. The change in z is denoted δz so that

$$z^{(1)} = z^{(0)} + \delta z. \tag{4-98}$$

4-5.1 AN APPROXIMATION OF THE PROBLEM OD

The basic idea in the approach to OD presented here is to construct an approximation of OD which can be solved to obtain an improvement δb in $b^{(0)}$. The approximate problem is obtained by making linear approximations to nonlinear functions in OD. Linear approximations to the changes in $f(z,b)$, $h(z,b)$, and $\phi_j(z,b)$ due to the small changes δb in $b^{(0)}$ and δz in $z^{(0)}$ are, by Taylor's Formula, Ref. 7, page 58

$$\delta f\left[z^{(0)}, b^{(0)}\right] = \frac{\partial f}{\partial z}\left[z^{(0)}, b^{(0)}\right]\delta z + \frac{\partial f}{\partial b}\left[z^{(0)}, b^{(0)}\right]\delta b \tag{4-99}$$

$$\delta h\left[z^{(0)}, b^{(0)}\right] = \frac{\partial h}{\partial z}\left[z^{(0)}, b^{(0)}\right]\delta z + \frac{\partial h}{\partial b}\left[z^{(0)}, b^{(0)}\right]\delta b \tag{4-100}$$

and

$$\delta\phi\left[z^{(0)}, b^{(0)}\right] = \frac{\partial\phi}{\partial z}\left[z^{(0)}, b^{(0)}\right]\delta z + \frac{\partial\phi}{\partial b}\left[z^{(0)}, b^{(0)}\right]\delta b. \tag{4-101}$$

In the development that follows, the arguments $[z^{(0)}, b^{(0)}]$ of all functions will be understood unless otherwise explicitly noted. The symbol δ in front of a quantity simply denotes the total differential of that quantity.

Since $h[z^{(0)}, b^{(0)}] = 0$ and $z^{(0)} + \delta z$ is to satisfy the equation $h[z^{(0)} + \delta z, b^{(0)} + \delta b] = 0$, the linearized version of this condition is simply

$$\frac{\partial h}{\partial z}\delta z + \frac{\partial h}{\partial b}\delta b = 0. \tag{4-102}$$

Eq. 4-102 is viewed as determining δz as a function of δb. It is clear that Eq. 4-102 can be solved for δz since the matrix $\partial h/\partial z$ has been assumed nonsingular.

Inequality constraints, Eq. 4-96, will be treated in an approximate manner. The method employed here is to require that if $\phi_j[z^{(0)}, b^{(0)}] > 0$, then

$$\delta\phi_j < \Delta\phi_j, \tag{4-103}$$

where $\Delta\phi_j$ is the required change in the value of ϕ_j due to the changes δz and δb in $z^{(0)}$ and $b^{(0)}$.

For convenience of notation, define the set of indices

$$A = \left\{ j : \phi_j\left[z^{(0)}, b^{(0)}\right] > 0 \right\}. \tag{4-104}$$

The set A (possibly empty) simply contains all the indices j of constraints that will be required to satisfy Eq. 4-103. To make maximum use of vector calculus notation, define the column matrix

$$\bar{\phi}(z,b) = \begin{bmatrix} \phi_j(z,b) \\ j \in A \end{bmatrix} \tag{4-105}$$

If the set A is empty, then $\bar{\phi}$ is defined as zero; i.e., all the constraint functions whose indices are in A are placed in a column matrix. In this way, the conditions, Eq. 4-103, may now be written.

$$\frac{\partial\bar{\phi}}{\partial z}\delta z + \frac{\partial\bar{\phi}}{\partial b}\delta b < \Delta\bar{\phi}, \tag{4-106}$$

where the column matrix $\Delta\bar{\phi}$ is defined as

$$\Delta\bar{\phi} = \begin{bmatrix} \Delta\phi_j \\ j \in A \end{bmatrix} \tag{4-107}$$

If A is empty, $\Delta\bar{\phi}$ is defined to be zero.

The object of the following analysis will be to choose δb so that $f[z^{(0)} + \delta z, b^{(0)} + \delta b]$ is as small as possible. If this nonlinear function of δz and δb is replaced by its Taylor approximation, the problem is to choose δz and δb to minimize

$$\delta f = \frac{\partial f}{\partial z}\delta z + \frac{\partial f}{\partial b}\delta b. \tag{4-108}$$

The entire argument up to this point has been based on the fact that $\|\delta b\|$ will be small. In order to insure that this is the case, it will be required that

$$\delta b^T W \delta b < \xi^2 \tag{4-109}$$

for ξ small and W a positive definite matrix. The matrix W will be used in particular problems to assign weights to the various components of δb. This is often necessary when the components of b represent different physical quantities that may be of different orders of magnitude. Usually W is diagonal.

To summarize the approximate problem, δb and δz are to be chosen to minimize

$$\frac{\partial f}{\partial z}\delta z + \frac{\partial f}{\partial b}\delta b \tag{4-110}$$

subject to the constraints

$$\frac{\partial h}{\partial z}\delta z + \frac{\partial h}{\partial b}\delta b = 0, \tag{4-111}$$

$$\frac{\partial \bar{\phi}}{\partial z}\delta z + \frac{\partial \bar{\phi}}{\partial b}\delta b < \Delta\bar{\phi} \tag{4-112}$$

and

$$\delta b^T W \delta b < \xi^2. \tag{4-113}$$

4-5.2 SOLUTION OF THE APPROXIMATE PROBLEM

Necessary conditions of Theorem 4-9 could now be applied directly to the approximate problem, Eqs. 4-110 through 4-113. If this course of action is followed, however, an explicit inverse of $\partial h/\partial z$ must be computed. Since the dimension n of this matrix is often quite high, this operation would be very costly. Instead of applying necessary conditions immediately, Eq. 4-111 will be used to eliminate the dependence of the remaining functions of the problem on δz. Necessary conditions may then be easily applied for the determination of δb.

The obvious method of eliminating dependence on δz is to solve Eq. 4-111 for δz as a function of δb. This, however, requires the inversion of the matrix $\partial h/\partial z$. The preceding approach of applying necessary conditions was scuttled for just this reason, so another method of eliminating δz must be found. Note that if the terms $(\partial f/\partial z)\,\delta z$ and $(\partial\bar{\phi}/\partial z)\,\delta z$ could be found in terms of δb, then dependence on δz would be eliminated. This is the approach that will be taken here and also in a later chapter on infinite dimensional problems.

Define the column matrix λ^J as the solution of

$$\frac{\partial h^T}{\partial z}\lambda^J = \frac{\partial f^T}{\partial z} \tag{4-114}$$

and the matrix $\lambda^{\bar{\phi}}$ as the solution of

$$\frac{\partial h^T}{\partial z}\lambda^{\bar{\phi}} = \frac{\partial \bar{\phi}^T}{\partial z}. \tag{4-115}$$

Note that $\lambda^{\bar{\phi}}$ is a matrix whose columns are solutions of

$$\frac{\partial h^T}{\partial z}\lambda^{\phi_j} = \frac{\partial \phi_j{}^T}{\partial z} \tag{4-116}$$

for $j \in A$. Note that Eqs. 4-114 and 4-115 require the repeated solution of equations with the same matrix on the left and different right-hand sides. There are efficient computation codes which can construct all the solutions simultaneously.

To see how these newly defined matrices are helpful, compute the transpose of both sides of Eqs. 4-114 and 4-115 and multiply through on the right by δz to obtain

$$\lambda^{J\,T}\frac{\partial h}{\partial z}\delta z = \frac{\partial f}{\partial z}\delta z \tag{4-117}$$

and

$$\lambda^{\phi T} \frac{\partial h}{\partial z} \delta z = \frac{\partial \phi}{\partial z} \delta z. \tag{4-118}$$

Note that the terms on the right side of these equations are exactly the ones which are to be eliminated from Eqs. 4-110 and 4-112. Further, the term $(\partial h/\partial z)\delta z$ that appears in both left-hand sides can be obtained from Eq. 4-111 as

$$\frac{\partial h}{\partial z}\delta z = -\frac{\partial h}{\partial b}\delta b.$$

Using this relation, Eqs. 4-117 and 4-118 become

$$-\lambda^{J T}\frac{\partial h}{\partial b}\delta b = \frac{\partial f}{\partial z}\delta z$$

and

$$-\lambda^{\phi T}\frac{\partial h}{\partial b}\delta b = \frac{\partial \phi}{\partial z}\delta z.$$

Substituting these relations into Eqs. 4-110 and 4-112, the approximate problem becomes: δb is to be chosen to minimize

$$\ell^{J T}\delta b \tag{4-119}$$

subject to the constraints

$$\ell^{\phi T}\delta b < \Delta\bar{\phi} \tag{4-120}$$

$$\delta b^T W \delta b < \xi^2 \tag{4-121}$$

where

$$\ell^J = \frac{\partial f^T}{\partial b} - \frac{\partial h^T}{\partial b}\lambda^J \tag{4-122}$$

and

$$\ell^{\phi} = \frac{\partial \phi^T}{\partial b} - \frac{\partial h^T}{\partial b}\lambda^{\phi}, \text{ if } A \text{ is not empty}$$

$$\ell^{\phi} = 0, \text{ if } A \text{ is empty.} \tag{4-123}$$

It should be noted that if the limitation, Eq. 4-121, on the size of $\|\delta b\|$ is not enforced, then the problem, Eqs. 4-119 and 4-120, is just a linear programming problem that may be solved by well-established techniques of linear programming. This technique is similar to that used in Zoutendijk's method of feasible directions (Ref. 8). For a discussion of this method the reader is referred to the literature.

The necessary conditions of Theorem 4-9 may now be applied to this reduced problem. In order to apply the theorem and in later calculations, it is required that the matrix $\ell^{\phi T}$ have full row rank; i.e., that the rows of $\ell^{\phi T}$ (columns of ℓ^{ϕ}) are linearly independent. Further, for use of the theorem it is required that the column vector $W\delta b$ be linearly independent of the columns of ℓ^{ϕ}. It may be noted that these assumptions require that there can be no more than $k-1$ constraint functions which are zero or positive at any iteration. This is true since the matrix ℓ^{ϕ} has only k rows and since its columns must be linearly independent of ℓ^{ϕ}, there can be at most $k-1$ remaining linearly independent columns. These assumptions are reasonable from a physical point of view. If ℓ^{ϕ} had rank k then the equation

$$\ell^{\phi}\delta b = \Delta\phi$$

would uniquely determine δb, and there would be no optimization problem.

The constraints, Eqs. 4-120 and 4-121, will be treated differently, so different multiplier notation in Theorem 4-9 will be used for each. First, define

$$\bar{H} = \ell^{J^T} \delta b + \bar{\mu}^T \ell^{\phi^T} \delta b + \nu \delta b^T W \delta b.$$

Theorem 4-9 requires that

$$\frac{\partial H}{\partial \delta b} = 0 = \ell^{J^T} + \bar{\mu}^T \ell^{\phi^T} + 2\nu \delta b^T W \quad (4\text{-}124)$$

where $\bar{\mu}_i \geq 0$ and $\nu \geq 0$

$$\bar{\mu}_i(\ell^{\phi^i} \delta b - \Delta\phi_i) = 0, i \in A \quad (4\text{-}125)$$

and

$$\nu(\delta b^T W \delta b - \xi^2) = 0. \quad (4\text{-}126)$$

At this point, a computational difficulty arises. It is difficult to determine δb from Eqs. 4-124, 4-125, and 4-126 since it is not known which of the constraints, Eqs. 4-120 and 4-121, will be equalities and which will be strict inequalities. The question is, "Which of the inequalities, Eq. 4-120 or Eq. 4-121, will become strict inequalities?" This can be interpreted geometrically as a question of leaving the boundary and going into the interior of the constraint set defined by Eqs. 4-120 and 4-121. It has been the experience with this technique that once a constraint, say $\phi_i(z, b)$, becomes zero, then for several small steps δb it will remain zero. This observation has led to the following computational procedure. First, all constraints, Eqs. 4-120 and 4-121, will be assumed equalities and δb is determined using Eqs. 4-124, 4-120, and 4-121. Then the algebraic signs of the $\bar{\mu}_i$ and ν are checked. If they are all non-negative, then this is the desired solution of the problem. If, on the other hand, some $\bar{\mu}_i$ or ν are negative, then the constraints corresponding to these multipliers are removed from Eq. 4-120 or Eq. 4-121 and the problem is again solved with the reduced number of constraints.

In any method of solution of the approximate problem, no information is gained if $\nu = 0$. Therefore, in the following $\nu > 0$ will be assumed.

Solving Eq. 4-124 for δb,

$$\delta b = -\frac{1}{2\nu} W^{-1}(\ell^J + \ell^{\phi} \bar{\mu}). \quad (4\text{-}127)$$

It is now assumed (to be checked later) that Eq. 4-120 is an equality. Substituting for δb from Eq. 4-127 into the equality Eq. 4-120,

$$-\frac{1}{2\nu} \ell^{\phi^T} W^{-1} (\ell^J + \ell^{\phi} \bar{\mu}) = \Delta\bar{\phi}.$$

Rewriting this equation,

$$\ell^{\phi^T} W^{-1} \ell^{\phi} \bar{\mu} = -\ell^{\phi^T} W^{-1} \ell^J - 2\nu \Delta\bar{\phi}.$$

Since ℓ^{ϕ^T} is required to have full row rank and W^{-1} is nonsingular, the matrix

$$M_{\phi\phi} = \begin{cases} 1 & \text{, if } A \text{ is empty} \\ \ell^{\phi^T} W^{-1} \ell^{\phi} & \text{, if } A \text{ is not empty} \end{cases} \quad (4\text{-}128)$$

is nonsingular. Therefore,

$$\bar{\mu} = -M_{\phi\phi}^{-1} \left(\ell^{\phi^T} W^{-1} \ell^J + 2\nu \Delta\bar{\phi}\right). \quad (4\text{-}129)$$

Note that in the unconstrained case when A is empty, $\mu = 0$ since $\ell^{\phi^T} = 0$ and $\Delta\bar{\phi} = 0$.

Substituting from Eq. 4-129 into Eq. 4-127

$$\delta b = -\frac{1}{2\nu} W^{-1}\left(I - \ell^{\phi} M_{\phi\phi}^{-1} \ell^{\phi^T} W^{-1}\right)\ell^J$$

$$+ W^{-1} \ell^{\phi} M_{\phi\phi}^{-1} \Delta\bar{\phi} \quad (4\text{-}130)$$

This expression for δb could now be substituted into $\delta b^T W \delta b = \xi^2$ to solve for ν. However, in practice it seems just as realistic to choose $\nu > 0$ in an iterative process as to choose ξ. Once $\nu > 0$ has been chosen $\bar{\mu}$ may be evaluated in Eq. 4-129. If any components are negative, the corresponding elements of ϕ are removed and δb is calculated using the new ψ matrix.

To aid in interpreting the meaning of terms in Eq. 4-130 for δb, define

$$\delta b^1 = W^{-1}\left(I - \ell^{\phi} M_{\phi\phi}^{-1} \ell^{\phi^T} W^{-1}\right)\ell^J \quad (4\text{-}131)$$

and

$$\delta b^2 = W^{-1} \ell^{\phi} M_{\phi\phi}^{-1} \Delta\phi \quad (4\text{-}132)$$

In this notation,

$$\delta b = -\frac{1}{2\nu}\delta b^1 + \delta b^2. \quad (4\text{-}133)$$

The vector δb^1 may be interpreted as a constrained gradient with $1/2\nu$ taken as a step size. The matrix which multiplies ℓ^J in Eq. 4-131 essentially projects the gradient ℓ^J of the cost function onto a tangent plane to the constraint set. The term δb^2 serves to drive any errors in constraint functions to zero. These interpretations are supported by Theorem 4-16.

Theorem 4-16: The vectors δb^1 and δb^2 of Eqs. 4-131 and 4-132 have the following properties:

1. $\delta b^{2^T} W \delta b^1 = 0$
2. $\ell^{\phi^T} \delta b^1 = 0$
3. $\ell^{\phi^T} \delta b^2 = \Delta\phi$
4. $-\ell^{J^T} \delta b^1 < 0.$

An obvious check on convergence is to monitor δb and the associated reduction in f, δf. When small δb occur and essentially no improvement is made in f, the process is terminated. This test, however, leaves a great deal to be desired since the choice of ν can yield very small steps δb and falsely lead the designer to believe that the iterative process is converging.

A much better test is to monitor the constrained gradient δb^1. Since in an unconstrained problem the gradient must approach zero at a minimum, one might expect that once $\Delta\phi = 0$, the constrained gradient δb^1 should approach zero. The real quantity $\|\delta b^1\|$ could then serve as a convergence check. Theorem 4-17 makes these ideas more rigorous.

Theorem 4-17: Let $f(z, b)$, $h(z,b)$, and $\phi(z, b)$ be continuously differentiable functions. If the sequences $\{b^{(j)}\}$ and $\{z^{(j)}\}$ generated by the above algorithm converge to the solution, $\bar{z}$, $\bar{b}$ of the problem OD and if $\phi = 0$ for all sufficiently large j, then it is necessary that δb^1 approaches zero as j approaches ∞.

4-5.3 STEEPEST DESCENT ALGORITHM

The iterative procedure developed in this paragraph may be summarized as follows:

Step 1. Make an engineering estimate of the optimum design variable, $b^{(0)}$.

Step 2. In the jth iteration, $j \geq 0$, solve Eq. 4-95 for $z^{(j)}$ corresponding to $b^{(j)}$.

Step 3. Form the vector of constraint functions $\hat{\phi}$ in Eq. 4-105 and solve Eqs. 4-114 and 4-115 for λ^J and $\lambda^{\hat{\phi}}$.

Step 4. Compute ℓ^J and $\ell^{\hat{\phi}}$ in Eqs. 4-122 and 4-123.

Step 5. Choose $\Delta\hat{\phi}$ in Eq. 4-107.

Step 6. Compute $M_{\hat{\phi}\hat{\phi}}$ in Eq. 4-108.

Step 7. Compute δb^1 and δb^2 in Eqs. 4-131 and Eq. 4-132.

Step 8. Choose $\nu > 0$ and evaluate $\bar{\mu}$ in Eq. 4-129. If any components of $\bar{\mu}$ are negative, take the corresponding elements out of $\hat{\phi}$ and return to Step 3.

Step 9. Compute

$$b^{(j+1)} = b^{(j)} - \frac{1}{2\nu}\,\delta b^1 + \delta b^2 .$$

Step 10. If $|f[x^{(j+1)}] - f[x^{(j)}]|$ and $\|\delta b^1\|$ are sufficiently small, terminate. Otherwise, return to Step 2.

4-5.4 USE OF THE COMPUTATIONAL ALGORITHM

The algorithm presented in par. 4-5.3 will certainly not solve all optimization problems. It is presented primarily to guide the designer to the proper equations developed in par. 4-5 while he is solving a problem. Almost surely a complicated real-world optimal design problem will have some feature which is not explicitly contained in the general formulation OD. In order to utilize a steepest-descent philosophy similar to the one developed here, the designer should be familiar with the method of obtaining the given algorithm. In this way, problems with peculiar features often can be treated by altering the general algorithm slightly.

There are two steps in the algorithm of par. 4-5.3 which are not complete. They are Steps 8 and 10. In Step 8, a parameter ν is to be chosen, but no analytical method of choosing it is given. This is the classical difficulty with steepest-descent methods. They give a direction but, unfortunately, they do not allow analytical determination of a step size $(1/(2\nu)$ in this case).

A simple technique for choosing ν which has worked well in a number of problems is given here as a candidate scheme. Since it is the δb^1 component of δb which tends to reduce f, the step size determination will be based on δb^1. The basic idea is to choose ν in order to obtain a certain percentage reduction in f. Let Δf (a negative quantity) be the desired reduction in f for a single iteration (perhaps a 5% to 10% reduction). Since for $\Delta\hat{\phi} \approx 0$,

$$\delta f = -\ell^{J\,T}\frac{1}{2\nu}\,\delta b^1 . \tag{4-134}$$

ν is chosen as

$$\nu = -\frac{\ell^{J\,T}\delta b^1}{2\Delta f} . \tag{4-134}$$

In many problems ν has been chosen according to Eq. 4-134 on the first iteration and held constant throughout the iterative process. In other problems convergence properties were improved if ν is changed during the iterative process.

REFERENCES

1. A.V. Fiacco and G.P. McCormick, *Nonlinear Programming: Sequential Unconstrained Minimization Techniques*, John Wiley & Sons, New York, 1968.

2. J. Abadie, Ed., *Nonlinear Programming*, John Wiley & Sons, New York, 1967.

3. D.J. Wilde and C.S. Beightler, *Foundations of Optimization*, Prentice-Hall, Englewood Cliffs, New Jersey, 1967.

4. W. Rudin, *Principles of Mathematical Analysis*, McGraw-Hill, New York, 1953.

5. R. Courant, "Variational Methods for the Solution of Problems of Equilibrium and Vibrations", *Bull. Amer. Math. Soc.*, 49, 1943, pp. 1-23.

6. I.O. Melts, "Nonlinear Programming Methods for Optimizing Dynamical Systems in Function Space", *Automation and Remote Control*, No. 1, January 1968, pp. 68-73.

7. C. Goffman, *Calculus of Several Variables*, Harper and Row, New York, 1965.

8. G. Zoutendijk, *Methods of Feasible Directions*, Elsevier Publishing Company, Amsterdam, 1960.

CHAPTER 5

FINITE DIMENSIONAL OPTIMAL STRUCTURAL DESIGN

5-1 INTRODUCTION

Throughout this handbook, structural optimization problems are chosen to illustrate the use of the design methods developed. There are two principal reasons for using structural problems for illustration. First, there has been great emphasis on helicopter and man portability of materiel, which places a premium on structural weight. Illustrative of Army concern with lightweight structures is the theme of the 1970 Army Mechanics Conference, "Lightweight Structures" (Ref. 33).

A second key reason for highlighting structural optimization is its advanced state of development, relative to other areas of the mechanical engineering sciences such as dynamics of machinery and mechanisms. A few examples in these related areas are treated in this handbook, but development of computational techniques remains to be done. It is felt that if the reader develops a thorough understanding of structural optimization and computational techniques, he will be in a good position to address problems outside the realm of structures. The fact that the mathematics of structural analysis parallels that of related mechanical disciplines strengthens this feeling.

A cursory review of Army materiel needs convinces one that light weight is a requirement for a majority of weapon systems being developed by the Army. The high priority placed on air mobility as well as lightweight infantry equipment has presented weapon system designers with a major challenge. In the case of air mobility, minimum equipment weight is a necessary condition for maximum helicopter payload. In infantry applications equipment weight limits the soldier's firepower and mobility.

In seeking lightweight designs, one is tempted to simply use lightweight materials and lower safety factors. It becomes apparent, however, that structural weight reduction can significantly degrade system performance. For example, when the weight of an artillery piece is reduced by 30%, dynamic response due to firing the weapon becomes much more severe. In infantry weapons, the requirement of reduced weight has led designers to lighter weight operating mechanisms for individual weapons. In lightweight rifles, for example, bolts are much lighter than in previous weapons and hence are more sensitive to changes in friction due to dust and external particles than were the more massive bolts in the M14 and M1 Rifles. There are many examples, some of which will be discussed later in this handbook, of instances in which simply reducing weight of subsystems causes problems which did not occur in heavier designs.

The lightweight objective, then, requires that the developer take an overall system view and consider the interaction between weapon weight and performance of the weapon system. As is true in virtually every design problem in which the limits of technology are approached, the lightweight weapon design problem must be considered simultaneously with all other aspects of system design. It is not practical to expect, therefore, that one will find lightweight structural design specialists operating independently of designers concerned with other aspects of the weapon development. A technology is needed which will allow trade-offs concerning weapon-weight to be integrated into the overall weapon design process.

The objective in this treatment has been to formulate the minimum weight structural design problem with constraints realistically reflecting the performance requirements of the weapon system. A detailed formulation and solution of this structural design problem is presented in this Chapter as well as Chapters 7 and 9.

5-1.1 LIGHTWEIGHT VS STRUCTURAL PERFORMANCE TRADE-OFFS

Normally, achieving a lightweight structure requires a reduction in the amount of material used. The consequence is an increase in structural flexibility that causes increased deflections, decreased natural frequencies, and decreased buckling loads. Consequently, failure modes that were not previously critical, may now become limiting factors in design. For example, in gun supporting structures, increased deflection often reduces effectiveness of the weapon system by increasing dispersion. There are many ways in which such changes in structural performance can have an impact on overall system behavior.

The only effective approach to minimum weight structural design is to formulate the structural design problem to include constraints on performance which are dictated by functional use of the weapon system. As a result, the minimum weight design problem is often stated with explicit constraints on structural deflection, natural frequency, buckling load, and strength. A central part of the design problem, then, is representation of weapon system performance requirements that have an impact on structural design. It is often required that in doing structural design, dynamic weapon performance must be analyzed to assure that the proper constraints are included in the structural design problem.

5-1.2 WEAPON DEVELOPMENT PROBLEMS ASSOCIATED WITH LIGHTWEIGHT REQUIREMENTS

To further explore some of the trade-offs between lightweight and weapon system performance, several typical problems encountered in weapon development will be discussed in this paragraph. The discussion here is presented to highlight some typical problems, not necessarily to identify all lightweight structural design problems faced in weapon development.

5-1.2.1 AIRCRAFT ARMAMENT

Some of the most critical lightweight structural development problems in weaponry today are in the field of aircraft armament. This is due to the very high priority placed by the Army on improved air mobility and the need for minimum weight weapon systems to be carried by helicopters. The combination of lightweight structural requirements and the extreme environment under which the structure must perform in helicopter application, generates a very difficult class of minimum weight structural design problems. The weapon developer's interest in structural design for aircraft armament lies primarily in the area of weapon and weapon support structures.

The key structural requirement in this application is accurate aiming of an automatic weapon during firing. Dynamic response of the weapon support structure due to inputs from the weapon and from the airframe, which vibrates due to aerodynamic inputs, must be considered in the design problem. The most difficult feature of the minimum weight structural design problem for aircraft weapon applications is the variety of performance and failure constraints which must be treated in the design process. Constraints must generally be placed on stresses arising in the structure, angular deflection of the structure at the gun mount, and natural frequency of the supporting structure. These constraints generally appear in the form of inequalities. For example, stress is required to be less than or equal to the allowable stress for the material. This kind of constraint is very realistic, from an engineering point of view, but makes the solution of the optimal design problem rather difficult.

In addition to altering the geometry and distribution of material in the structure to obtain desirable performance, it is also possible to induce damping into the structure and to use active feedback control devices to reduce response. These two methods of reducing dynamic response will require additional weight on board the helicopter. There is a trade-off between design of the structure and design of other means of obtaining improved weapon system performance. These trade-offs, then, require that we treat the aircraft weapon design problem as a system problem, explicitly accounting for the interaction between structural behavior, damping, and active feedback control.

5-1.2.2 GUN BARREL DESIGN

A second area in which lightweight structural design is of critical importance is that of gun barrel development, particularly for infantry automatic weapon application. With a great deal of emphasis being placed on lightweight infantry weapons, the barrel is a natural component in which to seek weight reduction. This is particularly true for rapid fire weapons in which heavy barrels have traditionally been used to alleviate temperature problems. For a particular barrel configuration, decreased mass tends to cause elevated temperatures and stresses. To complicate matters, material strengths are highly temperature dependent, making stress constraints difficult to handle. Another potential problem, as one tends toward optimality in barrel design, is the possibility that material yield properties will become critically dependent upon strain rates and require their explicit inclusion in the design process.

Another problem, which can arise in reduced weight design, is barrel deflection with resulting reduction in weapon accuracy. Deflection constraints must, therefore, be considered.

The objective of the barrel design problem is to choose barrel dimensions and structural material to minimize barrel weight in the presence of constraints on dollar cost, temperature, stress, and perhaps strain rate. The optimal design problem must then include equations of state of stress and temperature as a function of time, both depending on the barrel design features.

5-1.2.3 TOWED ARTILLERY

The principal objective in towed artillery design is to provide support for a large-caliber tube that will, upon firing, transmit momentum to the earth without doing damage to the support structure and without undue dynamic response. The fundamentals of the design problem then lie in the field of mechanics and, in particular, are highly dependent upon the weight distribution within the artillery piece.

In traditional artillery design, the support structure is flexible but has been quite heavy and stiff in the past so that the flexibility of the structure was a higher order effect. Also, heavier carriages reduced the severity of the dynamic response problem due to their higher mass. Recent developments, such as the M102, 105 mm Howitzer, have resulted in a weapon that weighs approximately 3200 lb, as compared to the older M101 which weighed 4500 lb. As a result of the reduced weight, problems have arisen in providing a firm support for the artillery piece on soil. More recent design efforts, including the XM164, 105 mm Howitzer, and XM198, 155 mm Howitzer, have resulted in weapons which are considerably lighter than their predecessors. As a result of the reduced structural weight of the weapon, dynamic response in both of these weapons became critical and had to be treated as a key design constraint in development of the recoil mechanism. For a discussion of a particular problem, the reader is referred to the artillery design example of par. 8-5.

Although these are primarily mechanical system design problems, they have arisen due to the lightweight design criterion. For this reason, when one considers lightweight structural design he must be willing to fit his structural design problem into a larger system design program and clearly understand the interfaces arising between structural and other system performance characteristics.

5-1.2.4 OTHER WEAPON PROBLEMS

The example problems cited in par. 5-1.2 are meant to illustrate the essential features of some of the more complex lightweight structural design problems faced in weapon development. They are simplifications of the real problems but are difficult enough to illustrate the need for research in development of design methods. In view of the current emphasis within the Army on air mobility and lightweight systems, new design methods are required which are capable of solving these and many more lightweight design problems.

5-1.3 PLAN FOR TECHNIQUE DEVELOPMENT

The remainder of this chapter will be devoted to formulation and application of a method of structural optimization. As noted at the beginning of par. 5-1, an in-depth treatment of lightweight structural design provides insight into application of the general methods of Chapter 4.

For a comprehensive review of structural optimization through 1967, the reader is referred to Refs. 1 and 2. Several of the major classes of optimal structural design problems are outlined in Ref. 2. Some of the key papers which have appeared in the literature since 1967 are listed in Refs. 3 through 18.

5-2 ELEMENTS OF THE ELASTIC STRUCTURAL DESIGN PROBLEM

A class of optimal structural design problems in which the structure must remain elastic is treated in this paragraph. The objective of this paragraph is to show how the optimization methods of Chapter 4 can be used to solve realistic optimal design problems. No attempt is made here to present a complete theory of optimal structural design that is capable of solving all problems

The reader should note that, even for the class of problems considered here, it is not possible to blindly apply the techniques of Chapter 4. A certain amount of knowledge of structural analysis is required before a reasonable statement of the design problem and a method of solution can be obtained. Even more important, the structural designer needs to have a thorough knowledge of the optimization methods of Chapter 4 and their development. As will be seen, in some cases it is required that parts of the design problem be interpreted in light of the derivation of the optimization method. In this way the method may be adapted for solution of a particular class of problems.

5-2.1 THE OPTIMALITY CRITERION

The meaning of optimal or best must be clearly established in each problem of interest. In order to have a problem which may be solved by the previously developed optimization methods, a real valued measure of the cost of the structure (value of the structure) must be chosen. Such measures as dollar cost of the structure, weight of the structure, or dynamic response of the structure may be chosen.

Along with the choice of a cost function, the parameters, or design variables, that represent all design alternatives must be chosen. These parameters will often be dimensions of structural members, area of member cross sections, or locations of joints in the structure. In keeping with the notation of the preceding chapter, these design variables will be denoted as b_i, $i = 1, \ldots, m$. For convenience of notation, these variables will be put in the vector form $b = [b_1, \ldots, b_m]^T$.

Invariably, the behavior of the structure under load will have to be considered in the design problem. The response of the structure may include quantities such as stress, displacement, buckling loads, and natural frequency. The collection of all variables required to describe this response due to applied load will be denoted by the state variable vector z. The manner in which z is related to the design variables and applied loads will be discussed in some detail later in this paragraph.

The cost of the structure must now be described as a real valued function of the design and behavior variables. In keeping with the preceding notation this function will be denoted as

$$J = f(z, b, \zeta) \qquad (5\text{-}1)$$

where ζ is one or more eigenvalues such as buckling load and natural frequency. Before a meaningful discussion of treatment of the structural design problem may be given, the behavior of the structure due to loads and constraints on that behavior must be analyzed.

5-2.2 STRESS AND DISPLACEMENT DUE TO STATIC LOADING

It is assumed for now that the structure of interest is either made up of a finite number of distinct interconnected members or that large continuous members in the structure have been approximated by a finite number of elements as in finite element techniques. Further, it is assumed that the entire structure is described by a vector design variable b.

Let stresses at critical points in the structure be denoted $z_1, \ldots, z_r$ and displacements required for the analysis and design of the structure be denoted $z_{r+1}, \ldots, z_n$. The behavior of the structure due to any given load may then be specified by the vector state variable $z = [z_1, \ldots, z_n]^T$. Attention will be restricted here and in the remainder of this chapter to structures which obey Hook's law, i.e., stress and displacement are determined by linear equations. It is clear, however, that the design variables play a large part in the response of the structure to loads. The dependence on the design variables enters these linear equations through the coefficients. The equations for z will be denoted

$$A(b)z = P \tag{5-2}$$

where P is a matrix of loads and

$$A(b) = [a_{ij}(b)]_{n \times n} \tag{5-3}$$

is a matrix whose elements depend on the design variables.

In this formulation of the problem, z and P may be generalized state and load variables. Eq. 5-2 may be obtained through direct application of equilibrium and compatibility conditions or through application of a variational criterion for equilibrium. In today's structural analysis technology, Eqs. 5-2 are very likely to be obtained by finite element methods (Refs. 19, 20). If the structural analysis problem is properly formulated, the matrix $A(b)$ is nonsingular and z may be obtained by solving Eq. 5-2. It is assumed that the elements of the matrix $A(b)$ are differentiable with respect to b.

In most real-world structural design problems the structure is required to carry a whole family of loads that occur at different times in the life of the structure. The treatment here will be limited to a finite number of loads, denoted P^i, $i = 1, \ldots, s$. Associated with each load is a state z^i determined by Eq. 5-2.

Constraints on behavior of the structure due to each of the applied loads P may include bounds on stresses and displacements. These constraints can generally be written in the form

$$\phi(z, b, \zeta) < 0 \tag{5-4}$$

where $\phi(z,b,\zeta) = [\phi_1(z,b,\zeta), \ldots, \phi_t(z,b,\zeta)]^T$. The inequality constraints, Eq. 5-4, are required to be satisfied for each of the states z^i due to different applied loads P^i.

It is clear that the Eqs. 5-2 and constraints, Eqs. 5-4, fit into the formulation of the finite dimensional optimal design problem of par. 4-5. Treatment of the restrictions imposed by Eq. 5-4, however, must be delayed until similar restrictions due to other behavior constraints are accounted for. The entire problem will be treated in par. 5-3.

5-2.3 NATURAL FREQUENCY AND BUCKLING

As pointed out in par. 5-1, the desire to obtain lightweight structures has led to resonance problems and, likewise, buckling problems. It is necessary, then, that a meaningful optimal design methodology be capable of enforcing constraints on eigenvalues associated with the system response. The sort of constraint considered here is

$$\zeta > \zeta_0 \tag{5-5}$$

where ζ is buckling load or natural frequency and ζ_0 is a lower bound on that eigenvalue. More general restrictions than those of Eq. 5-5 are included in the general constraint, Eq. 5-4.

Much as in Eq. 5-2, the equations of vibration or buckling may be written in the form

$$K(b)y = \zeta M(b)y \tag{5-6}$$

where $y = [y_1, \ldots, y_n]^T$ is an eigenvector which plays the role of a state variable,

$$K(b) = [k_{ij}(b)]_{n \times n} \tag{5-7}$$

is generally symmetric positive definite matrix, and

$$M(b) = [m_{ij}(b)]_{n \times n} \tag{5-8}$$

is generally also a symmetric positive definite matrix. Eq. 5-6 is often obtained through a finite element formulation of the structural analysis problem (Refs. 19, 20).

There are many methods for obtaining the eigenvalue and associated eigenvector in Eq. 5-6. The first method requires that the inverse of $K(b)$ be computed. Multiplying through Eq. 5-6 by $K^{-1}(b)$,

$$K^{-1}(b)M(b)y = \frac{1}{\zeta}\, y \tag{5-9}$$

This problem is now in standard form and the largest eigenvalue of $K^{-1}(b)M(b)$ is sought. The power method of obtaining this eigenvalue is quite effective (Ref. 21). It is particularly effective when a good estimate of the eigenvector is available. In the iterative design technique, a good estimate is generally available from the previous iteration. The power method is, therefore, well suited for use in iterative techniques. This method does have the severe disadvantage that $K^{-1}(b)$ must be computed for each new b.

A different method of finding the smallest eigenvalue and associated eigenvector of Eq. 5-6 without computing $K^{-1}(b)$ is based on the Rayleigh quotient as discussed in par. 2-8 and Ref. 23. The smallest eigenvalue of Eq. 5-6 is obtained by choosing a normalized vector y which minimizes the quotient $y^T K(b)y/[y^T M(b)y]$. The minimum value of this quotient is the smallest eigenvalue. A direct method of minimizing the Rayleigh quotient is discussed in par. 2-8.

5-2.4 METHOD OF SOLUTION

In the preceding formulation of the optimal design problem, the cost functions and constraints associated with stress and displacement can be put into the format of the problem treated in par. 4-5. The constraints associated with natural frequency and buckling, however, are not of exactly the same form. One difficulty is that the coefficient matrix for the eigenvector y, $K(b) - \zeta M(b)$, must be singular at the solution. This clearly contradicts the assumption in par. 4-5 that the state equations uniquely determine the state variable.

This situation is a direct result of Murphy's law "if anything can go wrong it will". Actually, it is not realistic to expect that a mathematical formulation of the kind presented in par. 4-5 should contain all real-world design problems. Already, an important problem has been encountered which requires an understanding of the development of par. 4-5 in order to include the new problem in the steepest-descent algorithm. The eigenvalue problem, fortunately, can be treated very nicely by the steepest-descent technique. Development of the method will be done in par. 5-3.

5-3 STEEPEST DESCENT PROGRAMMING FOR OPTIMAL STRUCTURAL DESIGN

In order to obtain a steepest-descent algorithm for the design problem with constraints on eigenvalues, it is necessary to go back into the derivation of the algorithm of par. 4-5. The major effort required here will be the linearization of the structural design problem to obtain an approximate problem of the kind described by Eqs. 4-119 through 4-121.

5-3.1 LINEARIZED COST AND CONSTRAINT FUNCTIONS

Since the cost and constraint functions depend on z, b, and ζ, the first order perturbation in these functions due to small changes δz, δb, and $\delta\zeta$ in z, b, and ζ is

$$\delta f = \frac{\partial f}{\partial z}\delta z + \frac{\partial f}{\partial b}\delta b + \frac{\partial f}{\partial b}\delta\zeta \tag{5-10}$$

and

$$\delta\bar{\phi} = \frac{\partial\bar{\phi}}{\partial z}\delta z + \frac{\partial\bar{\phi}}{\partial b}\delta b + \frac{\partial\bar{\phi}}{\partial\zeta}\delta\zeta\,. \tag{5-11}$$

The problem of writing the perturbed cost and constraint functions explicitly in terms of δb now reduces to obtaining explicit expressions for the terms involving δz and $\delta\zeta$.

From Eqs. 4-117 and 4-118, and the perturbed state equation we obtain, just as Eq. 4-119,

$$\frac{\partial f}{\partial z}\delta z = -\lambda^{J^T}\frac{\partial}{\partial b}[A(b)z]\delta b \tag{5-12}$$

and

$$\frac{\partial\bar{\phi}}{\partial z}\delta z = -\lambda^{\bar{\phi}}\frac{\partial}{\partial b}[A(b)z]\delta b \tag{5-13}$$

where λ^J and $\lambda^{\bar{\phi}}$ are determined by

$$A\lambda^J = \frac{\partial f^T}{\partial z} \tag{5-14}$$

and

$$A\lambda^{\phi_j} = \frac{\partial\phi_j^T}{\partial z} \tag{5-15}$$

for each $\phi_j > 0$. All this follows since $h(z,b)$ in the general formulation is simply $A(b)z$ in the present problem so

$$\frac{\partial h}{\partial b} = \frac{\partial}{\partial b}[A(b)z]$$

and

$$\frac{\partial h}{\partial z} = A$$

with A symmetric so $A^T = A$.

Thus, the explicit dependence of Eqs. 5-10 and 5-11 on δz can be easily eliminated. It remains to determine $\delta\zeta$ in terms of δb. This problem has been addressed in a completely rigorous manner by Kato (Ref. 23). Explicit expressions are given there under quite restrictive hypotheses. A formal development will be given here which obtains the same result.

It is assumed that the eigenvalues and eigenvectors of

$$K(b)y = \zeta M(b)y \tag{5-16}$$

depend continuously on b and further, that to first order, the following perturbation equation is accurate

$$K(b)\delta y + \frac{\partial}{\partial b}[K(b)y]\delta b = \delta\zeta M(b)y + \frac{\partial}{\partial b}[M(b)y]\delta b + \zeta M(b)\delta y \tag{5-17}$$

where y and ζ satisfy Eq. 5-16.

If $K(b)$ or $M(b)$ is not symmetric, it is necessary to solve the adjoint eigenvalue problem.

$$K^T(b)\bar{y} = \zeta M^T(b)\bar{y} \tag{5-18}$$

that has the same eigenvalue ζ as Eq. 5-16 but a different eigenvector $\bar{y}$. Rearranging and premultiplying by $\bar{y}^T$ this is

$$\bar{y}^T[K(b) - \zeta M(b)]\,\delta y + \bar{y}^T \frac{\partial}{\partial b}[K(b)y]\,\delta b$$

$$-\bar{y}^T\zeta \frac{\partial}{\partial b}[M(b)y]\,\delta b$$

$$= \bar{y}^T \delta\zeta M(b)y$$

Since the first term is a scalar,

$$\bar{y}^T[K(b) - \zeta M(b)]\,\delta y = \delta y^T[K^T(b) - \zeta M^T(b)]\,\bar{y}.$$

Since $\bar{y}$ is an eigenvector of Eq. 5-18,

$$[K^T(b) - \zeta M^T(b)]\,\bar{y} = 0$$

and this equation becomes

$$\left\{\bar{y}^T \frac{\partial}{\partial b}[K(b)y] - \zeta\bar{y}^T \frac{\partial}{\partial b}[M(b)y]\right\}\delta b$$

$$= \delta\zeta\bar{y}^T M(b)y.$$

Assuming $\bar{y}^T M(b)y \neq 0$ which will generally be the case,

$$\delta\zeta = \left\{\bar{y}^T \frac{\partial}{\partial b}[K(b)y] - \zeta\bar{y}^T \frac{\partial}{\partial b}[M(b)y]\right\}\delta b \Big/ [\bar{y}^T M(b)y]. \tag{5-19}$$

Derivation of the perturbation formula, Eq. 5-19, has been strictly formal. The assumption that Eq. 5-17 holds is highly questionable from an operator theoretic point of view. Under reasonable assumptions on the finite dimensional eigenvalue problem treated here however, Eq. 5-19 is shown to hold (Ref. 23); i.e., even though the justification given here is not mathematically rigorous, the result, Eq. 5-19, holds for a large class of problems.

Defining

$$\ell^J = \frac{\partial f^T}{\partial b} - \left\{\frac{\partial}{\partial b}[A(b)z]\right\}^T \lambda^J + \left\{\frac{\partial f}{\partial \zeta}\Big/[\bar{y}^T M(b)y]\right\} \times \left(\left\{\frac{\partial}{\partial b}[K(b)y]\right\}^T \bar{y} - \zeta\left\{\frac{\partial}{\partial b}[M(b)y]\right\}^T \bar{y}\right) \tag{5-20}$$

and

$$\ell^\phi = \begin{cases} \dfrac{\partial \phi^T}{\partial b} - \left\{\dfrac{\partial}{\partial b}[A(b)z]\right\}^T \lambda^{\phi T} + \left(\left\{\dfrac{\partial}{\partial b}[K(b)y]\right\}^T \bar{y} - \zeta\left\{\dfrac{\partial}{\partial b}[M(b)y]\right\}^T \bar{y}\right) \times \dfrac{\partial \phi^T}{\partial b}\Big/[\bar{y}^T M(b)y], \\ \text{or} \\ 0, \text{ if } \phi \text{ is empty.} \end{cases} \tag{5-21}$$

Eqs. 5-10 and 5-11 become

$$\delta J = \ell^{J\,T}\,\delta b \tag{5-22}$$

and

$$\delta\phi = \ell^{\phi\,T}\,\delta b. \tag{5-23}$$

The linearized problem is now to minimize δJ, Eq. 5-22, subject to constraints

$$\ell^{\bar{\phi}^T} \delta b < \Delta\bar{\phi}$$

where $\Delta\bar{\phi}$ is the desired correction in constraint error and

$$\delta b^T W \delta b < \xi^2$$

where W is positive definite and ξ is small. This is precisely the same problem in par. 4-5, Eqs. 4-119 through 4-121, so the theoretical results and steepest-descent algorithm of that paragraph apply with proper interpretation.

5-3.2 STEEPEST DESCENT ALGORITHM FOR OPTIMAL STRUCTURAL DESIGN

Step 1. Make an engineering estimate of the optimum design variable $b^{(0)}$.

Step 2. For $j = 0, 1, \ldots,$ solve Eq. 5-2 for $z^{(j)}$, Eq. 5-6 for $y^{(j)}$ and $\zeta^{(j)}$, and Eq. 5-9 for $\bar{y}$ (if $k(b)$ or $M(b)$ is not symmetric) with $b = b^{(j)}$.

Step 3. Form $\bar{\phi}$ as in Eq. 4-105. Solve Eqs. 5-14 and 5-15 for λ^J and $\lambda^{\bar{\phi}}$.

Step 4. Compute ℓ^J and $\ell^{\bar{\phi}}$ in Eqs. 5-20 and 5-21.

Step 5. Choose $\Delta\bar{\phi}$ as the desired reduction in constraint error.

Step 6. Compute

$$M_{\phi\phi} = \begin{cases} 1, \text{ if } \bar{\phi} \text{ is empty} \\ \ell^{\bar{\phi}^T} W^{-1} \ell^{\bar{\phi}}, \text{ elsewhere} \end{cases}$$

Step 7. Choose $\nu > 0$ and calculate $\mu = -M_{\phi\phi}^{-1}(\ell^{\bar{\phi}^T} W^{-1} \ell^J + 2\nu\Delta\bar{\phi})$. If any component of μ is negative, remove the corresponding row from $\bar{\phi}$ and return to Step 3.

Step 8. Compute

$$\delta b^1 = W^{-1}(I - \ell^{\bar{\phi}} M_{\phi\phi}^{-1} \ell^{\bar{\phi}^T} W^{-1})\ell^J$$

and

$$\delta b^2 = W^{-1} \ell^{\bar{\phi}} M_{\phi\phi}^{-1} \Delta\bar{\phi}$$

and form

$$\delta b = -\frac{1}{2\nu}\delta b^1 + \delta b^2.$$

Step 9. Compute

$$b^{(j+1)} = b^{(j)} + \delta b.$$

Step 10. If all constraints are satisfied and δb^1 is sufficiently small, terminate. Otherwise, return to Step 2 and continue the process.

All the properties of δb^1 and δb^2 derived in par. 4-5.2 hold in this case. Further, the discussion of that paragraph regarding such things as choosing ν also hold. The reader should refer to that paragraph for detailed discussions.

5-3.3 COMPUTATIONAL CONSIDERATIONS

Several comments on the computational art used in solution of these problems are in order. First, if a feasible design was chosen initially, large steps could be taken until one or more constraints were violated, at which time the step size was reduced. Second, it was noted that as the optimum was approached, oscillation occurred. By monitoring the dot product, $\delta b^{(j)T} \cdot \delta b^{(j-1)}$, oscillations were sensed when negative values of the dot product occurred. Thus, step size, $1/(2\nu)$, was divided by two when negative values of the dot product occurred on two successive iterations. Finally, the most effective method of adjusting step size was to monitor successive reductions in cost function after feasibility had occurred. Once insignificant reductions occurred, the step size was reduced to obtain finer convergence.

The Power method used to compute the smallest eigenvalue performs quite well. At every iteration, the starting value for the eigenvector is taken from the previous iteration which manifested a very rapid rate of convergence. An accuracy of 0.1% in each component of the eigenvector was used to compute the new eigenvector. The stiffness matrix for the structure was inverted by the Gauss-Jordan elimination procedure.

Another comment that is appropriate here concerns the sign check on the Lagrange multiplier vector μ, called for in Step 7 of the computational algorithm (par. 5-3.2). The algebraic sign of each component of the Langrange multiple vector μ was checked at each iteration. If some of the components were negative, then the matrix ℓ^3 and the vector $\Delta\phi$ were adjusted accordingly. This procedure is particularly useful whenever there were redundant constraint violations. In some cases, the number of constraints violated is more than the number of design variables of the problem, yielding a singular matrix coefficient of μ. In such cases numerical noise yielded a solution such that some of the components of the vector μ were always negative, indicating that the corresponding constraints would be strictly satisfied in the next iteration. In numerical examples, the number of constraints with positive components of μ was always less than or equal to the number of design variables of the problem. This procedure of adjusting the constraint set has worked very well and has minimized the possibility of divergence of the algorithm.

The method presented is relatively automatic in the sense that, for the computer program developed, the input data given is the only pertinent design information required for solution of the problem. All the necessary matrices and their derivatives are automatically generated in the computer. Any person with a reasonable knowledge of FORTRAN language should be able to handle the programming without any difficulty. The method is developed to meet simultaneously displacement, strength, and frequency requirements on the structure. The technique, therefore, can be made user oriented.

5-4 OPTIMIZATION OF SPECIAL PURPOSE STRUCTURES

Several special purpose structural optimization problems are solved in this paragraph on an ad-hoc basis to illustrate the method of par. 5-3. Subsequent paragraphs will treat large scale problems in a more unified manner.

5-4.1 A MINIMUM WEIGHT COLUMN

A column is to be constructed by making its cross section piecewise uniform as shown in Fig. 5-1. The objective of the design problem is to choose the element areas so that the column will support a vertical load P_0 without buckling or yielding under compressive load. For the purpose of the present problem the geometric shape of each column element is fixed and symmetric about two

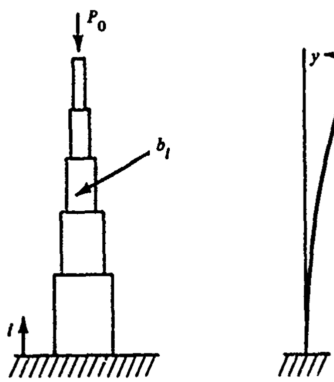


Figure 5-1. Column

orthogonal axes so that the cross-sectional area b_i of the ith element completely specifics the element. With this assumption, if α is the second moment of the cross section of unit area, then

$$I_i = \alpha b_i^2 \tag{5-24}$$

In this problem, weight of the column is to be minimized so that the cost function is

$$J = \gamma \sum_{i=1}^{k} b_i L_i \tag{5-25}$$

where γ is material density and L_i is the length of the ith element of the column.

There are two basic constraints that must be satisfied in this design problem. First, to insure that the buckling load P is not less than the applied load P_0, it is required that

$$\phi_{k+1} \equiv P_0 - P < 0. \tag{5-26}$$

Second, in order to insure that the column material does not yield under the applied load P_0, it is necessary that

$$\phi_i \equiv (P_0/b_i) - \sigma_{max} < 0, \quad i = 1, \dots, k \tag{5-27}$$

where σ_{max} is the allowable stress of the column material in compression.

In order to apply the optimization method of par. 5-3, the equations which determine the buckling load in terms of $b = [b_1, \dots, b_k]^T$ must be obtained. Using the generalized coordinates shown in Fig. 5-2 and Eq. B-4, Appendix B, the potential energy of the ith element under the buckling load P is

$$PE_i = \frac{1}{2} u^{i^T} K^i(b) u^i - P u^{i^T} D^i(b) u^i \tag{5-28}$$

where

$$u^i = [u_1^i, u_2^i, u_3^i, u_4^i]^T$$

is as shown in Fig. 5-2. The matrices $K^i(b)$ and $D^i(b)$ are from Eqs. B-4 and B-8, Appendix B

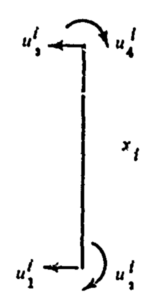


Figure 5-2. Column Element

$$K^i(b) = \frac{E\alpha b_i^2}{L_i^3} \begin{bmatrix} 12 & -6L_i & -12 & -6L_i \\ -6L_i & 4L_i^2 & 6L_i & 2L_i^2 \\ -12 & 6L_i & 12 & 2L_i \\ -6L_i & 2L_i^2 & 2L_i & 4L_i^2 \end{bmatrix} \tag{5-29}$$

and

$$D^i(b) = 2\begin{bmatrix} \frac{3}{5L_i} & -\frac{1}{20} & -\frac{3}{5L_i} & -\frac{1}{20} \\ -\frac{1}{20} & \frac{L_i}{15} & \frac{1}{20} & -\frac{L_i}{60} \\ -\frac{3}{5L_i} & \frac{1}{20} & \frac{3}{5L_i} & \frac{1}{20} \\ -\frac{1}{20} & -\frac{L_i}{60} & \frac{1}{20} & \frac{L_i}{15} \end{bmatrix} \qquad (5\text{-}30)$$

Summing the total potential energies of all the elements from Eq. 5-28 and defining a new variable

$$y = [y_1, y_2, ..., y_{2k}]^T$$

$$= [u_1^2, u_2^1, u_1^2, u_2^2, ..., u^k, u^k]^T$$

the total potential energy PE of the column may be written

$$PE = \frac{1}{2} y^T K(b)y - P\frac{1}{2} y^T D(b)y$$

where $K(b)$ and $D(b)$ are made up of elements of $K^i(b)$ and $D^i(b)$ and are symmetric. Applying the theorem of minimum total potential energy given in Appendix B, the governing equations of buckling are

$$K(b)y = P\,D(b)y. \qquad (5\text{-}31)$$

Eq. 5-31 is now in the form of Eq. 5-6, with proper interpretation of notation.

In order to implement the computational algorithm of par. 5-3, the following vectors are required:

$$\ell^{J^T} = \frac{\partial J}{\partial b} = [\gamma L_1, \gamma L_2, ..., \gamma L_k] \qquad (5\text{-}32)$$

$$\ell^{\phi_i^T} = \frac{\partial \phi_i}{\partial b} = [0, ..., 0, -P_0/b_i^2, 0, ..., 0],$$

$$i = 1, ..., k \qquad (5\text{-}33)$$

since ϕ_i does not depend on P, $i = 1,..., k$ and

$$\ell^{\phi_{k+1}} = \begin{cases} -\left(\left\{\frac{\partial}{\partial b}[K(b)y]\right\}^T y \right. \\ \left. -P\left\{\frac{\partial}{\partial b}[D(b)y]\right\}^T y\right) \Big/ (y^T D y), \\ \quad \text{if } \phi_{k+1} > 0 \\ [0], \text{ if } \phi_{k+1} < 0. \end{cases} \qquad (5\text{-}34)$$

The computations required in Eq. 5-34 are messy but they can be programmed for automatic computation.

All expressions required for direct application of the steepest descent algorithm of par. 5-3 are now available. Numerical results and profiles of optimum columns are shown in Tables 5-1 and 5-2, and Fig. 5-3. Numerical data for the example problems are $E = 3.0 \times 10^7$ psi, $\alpha = 0.079577$, $\sigma_{max} = 20{,}000$ psi, and $L = 10.0$ in. Computation in each case required approximately 0.1 sec per iteration

TABLE 5-1
COMPARISON OF UNIFORM AND OPTIMAL COLUMNS

P, lb	Volume of Optimal Column, in³	Volume of Uniform* Column, in³	Material Savings, %
500	0.806	0.923	12.7
1000	1.143	1.300	12.1
1500	1.411	1.600	11.8
2000	1.640	1.840	10.9
4000	2.412	2.600	7.2

**Lightest uniform column which will support load P.*

TABLE 5-2
CROSS-SECTIONAL AREAS OF OPTIMUM COLUMNS

Element No. i	P = 500 lb	P = 1000 lb	P = 1500 lb	P = 2000 lb	P = 4000 lb
1	0.1070	0.1499	0.1833	0.2106	0.2947
2	0.1055	0.1480	0.1809	0.2076	0.2875
3	0.1035	0.1442	0.1763	0.2023	0.2789
4	0.1000	0.1383	0.1691	0.1942	0.2683
5	0.0960	0.1303	0.1593	0.1831	0.2505
6	0.0831	0.1198	0.1464	0.1683	0.2302
7	0.0738	0.1064	0.1299	0.1493	0.2020
8	0.0623	0.0892	0.1088	0.1250	0.2000
9	0.0477	0.0668	0.0812	0.1000	0.2000
10	0.0267	0.0500	0.0750	0.1000	0.2000

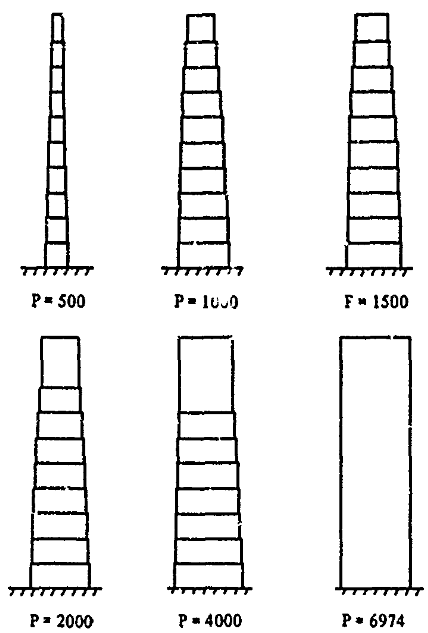


Figure 5-3. Profiles of Optimal Columns

and 15 iterations to converge on an IBM 360-65.

5-4.2 A MINIMUM WEIGHT VIBRATING BEAM

A beam is to be designed by piecing together uniform sections of beams as shown in Fig. 5-4. The objective is to choose the sections so that the beam is as light in weight as possible and still satisfies constraints on strength and natural frequency. Due to dynamic inputs to the beam, it is required that the natural frequency of the beam be above a given limit ω_0 to prevent oscillation problems.

As in the preceding column design problem, the cross-sectional geometry is chosen, but all dimensions of the cross section may be varied in the same proportion. Thus, if b_i denotes the area of the ith section, then the second moment of the cross-sectional area is

$$I_i = \alpha b_i^2 \tag{5-35}$$

where α is a constant of proportionality depending on the geometry of the cross section. The problem at hand is to minimize

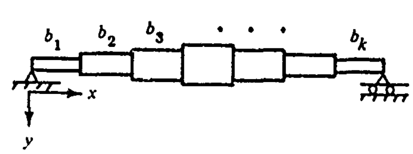


Figure 5-4. Stepped Beam

weight, so the cost function is

$$J = \rho \sum_{i=1}^{k} b_i L_i \tag{5-36}$$

where ρ is material density and L_i is the length of the ith section.

As a strength constraint, it is required that

$$\phi_i = b_0 - b_i < 0, \quad i = 1, \ldots, k \tag{5-37}$$

where $b_0 > 0$ is chosen so that the beam will support a lateral load. The constraint on natural frequency can be written as

$$\phi_{k+1} = \omega_0 - \omega < 0. \tag{5-38}$$

By neglecting compression of the beam, deformation of a typical element is shown in Fig. 5-5. By Appendix B, the kinetic energy

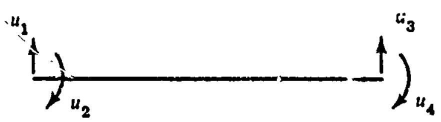


Figure 5-5. Typical Element

of an element is $\dot{u}^{i}M^{i}(b)\dot{u}^{i}/2$, where, from Eq. B-6

$$M^i(b) = \frac{\rho b_i L_i}{420}\begin{bmatrix} 156 & -22L_i & 54 & 13L_i \\ -22L_i & 4L_i^2 & -13L_i & -3L_i^2 \\ 54 & -13L_i & 156 & 22L_i \\ 13L_i & -3L_i^2 & 22L_i & 4L_i^2 \end{bmatrix}. \tag{5-39}$$

Likewise, the potential energy of the ith element is $u^{iT}K^i(b)u^i/2$, where, from Eq. B-4

$$K^i(b) = \frac{E}{L_i}\begin{bmatrix} 12I_i & -6I_iL_i & -12I_i & -6I_iL_i \\ -6I_iL_i & 4I_iL_i & 6I_iL_i & 2I_iL_i^2 \\ -12I_i & 6I_iL_i & 12I_i & 6I_iL_i \\ -6I_iL_i & 2I_iL_i^2 & 6I_iL_i & 4I_iL_i^2 \end{bmatrix}. \tag{5-40}$$

Forming a single vector y that contains all displacements and rotations for the beam, the total kinetic and potential energies are $\dot{y}^T M(b)\dot{y}/2$ and $y^T K(b)y/2$, respectively. Lagrange's equations, Eq. B-17, are then

$$M(b)\ddot{y}(t) + K(b)y(t) = 0 \tag{5-41}$$

For harmonic motion of the structure, $y(t) = y \sin \omega t$, where y is just a constant vector, t is time, and ω is natural frequency. Substituting into Eq. 5-41 and defining $\zeta = \omega^2$, the eigenvalue equation is

$$K(b)\, y = \zeta M(b) y \tag{5-42}$$

The problem of minimizing J of Eq. 5-36, subject to the constraints of Eqs. 5-37 and 5-38, and with state Eq. 5-42, is in the form of the general problem of par. 5-3. The steepest-descent computational algorithm of that paragraph can be applied directly to this problem.

As a numerical example, the beam problem was solved with the data $E = 3 \times 10^7$ psi, $L = 10$ in., $\alpha = 1.0$, and $\rho = 0.00208$ lb-sec^2/in^3. The computational algorithm required about 0.6 sec per iteration on an IBM 360-65 system and approximately 15 iterations to converge. Results for a range of natural frequencies are given in Table 5-3 and the profile of an optimum beam is shown in Fig. 5-6.

TABLE 5-3
COMPARISON OF OPTIMUM BEAMS

Frequency, rad/sec	Volume of Uniform Beam*, in^3	Optimum Volume, in^3	Material Savings, %
3600	0.935	0.897	4.06
4000	1.155	1.062	8.05
4400	1.397	1.259	9.74
4800	1.663	1.481	10.94
5200	1.951	1.727	11.48
5600	2.263	1.993	11.93
6000	2.598	2.283	12.12
10000	7.217	6.330	12.29

*Uniform beam of lowest volume having required natural frequency.

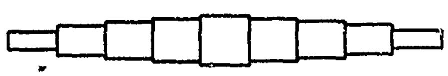


Figure 5-6. Profile of Optimum Beam

5-4.3 A MINIMUM WEIGHT PORTAL FRAME WITH A NATURAL FREQUENCY CONSTRAINT

A portal frame as shown in Fig. 5-7 is to be proportioned so that it weighs as little as possible and has its fundamental frequency at least as large as a specified frequency ω_0. Each member of the planar frame is formed from several uniform sections whose areas are to be determined as design variables. As in the preceding problems, the cross-sectional geometry is taken as fixed and all dimensions of cross sections varied proportionally. The second moment of the cross-sectional area about a centroidal axis is $I_i = \alpha b_i^2$ where b_i is the cross-sectional area of the ith element.

Neglecting strain energy due to axial deformation of the horizontal member, the element stiffness matrix from Appendix B is

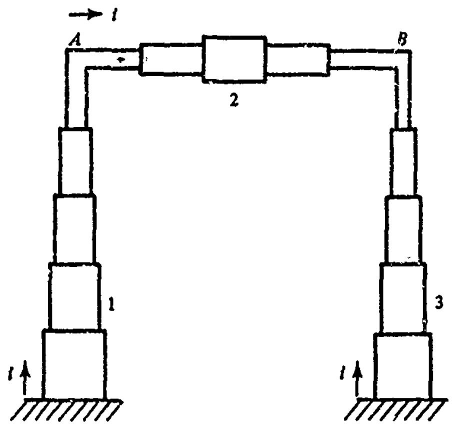


Figure 5-7. Portal Frame

$$K(b_i) = \frac{E\alpha b_i^2}{L_i^3}\begin{bmatrix} 12 & -6L_i & -12 & -6L_i \\ -6L_i & 4L_i^2 & 6L_i & 2L_i^2 \\ -12 & 6L_i & 12 & 6L_i \\ -6L_i & 2L_i^2 & 6L_i & 4L_i^2 \end{bmatrix} \tag{5-43}$$

where L_i is the length of the ith member and the element deformation variables are shown in Fig. 5-8. The potential energy PE of the ith element is

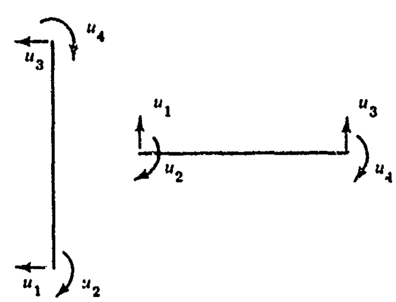


Figure 5-8. Typical Elements

$$PE_i = \frac{1}{2} u^{i^T} K(b_i) u^i \tag{5-44}$$

where $u^i = [u_1, u_2, u_3, u_4]^T$.

Likewise, from Appendix B the kinetic energy KE of the typical element is

$$KE_i = \frac{1}{2} \dot{u}^{i^T} M(b_i) \dot{u}^i$$

where $\dot{u}$ denotes time derivative of u and

$$M(b_i) = \frac{\rho b_i L_i}{420} \begin{bmatrix} 156 & -22L_i & 54 & 13L_i \\ -22L_i & 4L_i^2 & -13L_i & -3L_i^2 \\ 54 & -13L_i & 156 & 22L_i \\ 13L_i & -3L_i^2 & 22L_i & 4L_i^2 \end{bmatrix}. \tag{5-45}$$

Taking into account the lateral rigid body motion of Member 2, the total kinetic energy of the structure is

$$KE = \sum_i \frac{1}{2} \dot{u}^{i^T} M(b_i) \dot{u}^i + \frac{1}{2} \bar{M} \dot{u}_A^2 \tag{5-46}$$

where $\bar{M}$ is the mass of Member 2 and $\dot{u}_A$ is the horizontal velocity of point A.

Requiring harmonic motion with frequency ω, the displacement vector $y(t)$ made up of all displacements is

$$y(t) = y \sin \omega t$$

where y is a constant vector. Applying Lagrange's equations and eliminating time dependence yields

$$K(b)y = \zeta M(b)y \tag{5-47}$$

where y is the vector of all displacements and rotation, and $\zeta = \omega^2$. The matrices $K(b)$ and $M(b)$ are formed from element stiffness and mass matrices as outlined in Appendix B.

Eq. 5-47 is in exactly the form of Eq. 5-16 and the matrix for this problem is simply weight of the structure which is

$$J = \rho \sum_{i=1}^{k} b_i L_i \tag{5-48}$$

where ρ is density of the structural material.

The constraints imposed on the problem include lower limits on cross-sectional area

$$\phi_i = b_0 - b_i < 0, \quad i = 1, \ldots, k. \tag{5-49}$$

where $b_0 > 0$ and a lower limit on natural frequency

$$\phi_{k+1} = \zeta_0 - \zeta < 0 \tag{5-50}$$

where ζ_0 is the lowest allowable eigenvalue of Eq. 5-47, $\zeta_0 = \omega_0^2$.

The steepest-descent algorithm may now be applied directly. Data for the specific problems solved are given in Table 5-4. The results for an aluminum portal frame are given in Tables 5-5 and 5-6, with a typical profile shown in Fig. 5-9. The design variable b_i shows the distribution of material for a minimum weight frame whose frequency of vibration must be greater than or equal to a

TABLE 5-4
MATERIAL PROPERTIES FOR ALUMINUM

α, dimensionless	0.07058
ρ lb-sec^2/in.^{4}	2.616x10^4
E, lb/in.^{2}	10.3x10^6
I_0, in.^{4}	0.009826
L, in.	10.0

TABLE 5-5
COMPARISON OF UNIFORM AND OPTIMAL FRAMES FOR ALUMINUM

Frequency, rad/sec	Weight of Uniform Frame, lb	Weight of Optimal Frame, lb	Weight Reduction, %
2000	3.748	1.729	53.9
3000	8.434	2.562	69.6
4000	14.994	3.590	76.1
5000	23.428	4.688	80.0

TABLE 5-6
OPTIMAL DESIGN VARIABLE b_i FOR VIBRATING FRAME

b_i	ω, rad/sec 2000	3000	4000	5000
b_1	1.577	1.964	2.907	4.020
b_2	0.883	1.604	2.484	3.321
b_3	0.552	1.416	1.912	2.622
b_4	0.374	0.866	1.290	1.725
b_5	0.350	0.360	0.671	0.836
b_6	0.350	0.350	0.350	0.350

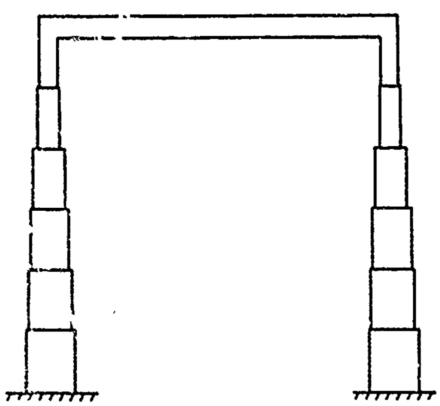


Figure 5-9. Optimum Portal Frame for ω = 3000 rad/sec

specified value. It can be seen from Table 5-5 that a significant material saving is possible in comparison to the portal frame with members of constant cross section.

The results for the design variables b_i are the same for Members 1 and 3, and Member 2 converges to the lower bound I_0, so only the results for Member 1 are reported. For all frequencies the values for the b_i for Member 2 are equal to 0.350.

5-4.4 A MINIMUM WEIGHT FRAME WITH MULTIPLE FAILURE CRITERIA

To illustrate the applicability of the steepest-descent method for the minimum weight design of structures with stress, buckling, and displacement constraints, an example of a statically loaded frame problem is presented. Fig. 5-10 shows the geometrical

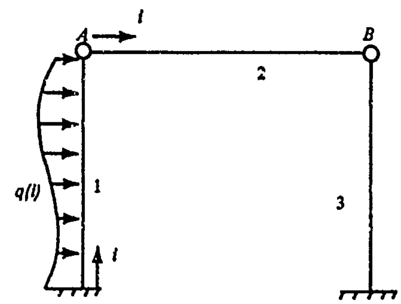


Figure 5-10 Frame With Side Loading

configuration of the frame that is considered. All members are assumed to be of the same length L. Member 1 is subjected to a lateral loading $q(l)$. Member 3 has a uniform cross-sectional area which is prescribed and will not be allowed to vary. The connections at points A and B are frictionless pins.

The finite element method is used to obtain the elastic response of the system for a given set of design variables, i.e., the cross-sectional areas of the elements. As in the preceding problem, the geometry of each cross section is the same with all dimensions of cross section varying proportionally. Thus, $I_i = \alpha b_i^2$ where b_i is the cross-sectional area of the ith element. The stiffness matrix $K(b_i)$ of a typical element, Fig. 5-11, can be written as in par. 5-4.3 with generalized displacements defined by

$$u^i = [u_1, u_2, u_3, u_4]^T.$$

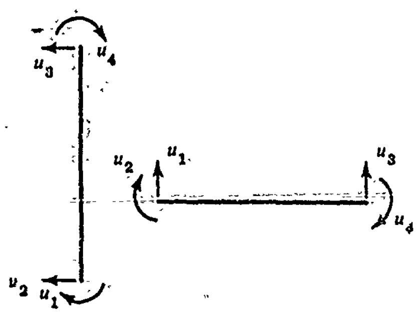


Figure 5-11. Typical Elements

From the fundamental beam theory, if R is the horizontal force transmitted from the Member 1 to 3, and assuming that Member 2 remains straight without buckling, then neglecting compression of Member 2, the deflection at A is $u_A = RL^3/(3EI_3)$. From the equilibrium conditions on the transverse forces and moments at the nodes of Member 1, the generalized displacement z, which is made up of the element displacements u_i can be evaluated from the following matrix equation

$$A(b)z = F \qquad (5\text{-}51)$$

where F is a vector load and $A(b)$ is a symmetric matrix. In a similar manner, if y is the displacement vector containing all element deflections associated with Member 2, the buckling load P can be determined by solving the eigenvalue problem

$$K(b)y = PD(b)y \qquad (5\text{-}52)$$

where the matrix D is derived from the shortening of Member 2 as in par. 5-4.1 and, as in the previous problem, $K(b)$ is a stiffness matrix.

The cost function to be minimized in this problem is the structural weight of Members 1 and 2 which is simply

$$J = \gamma \sum_{i=1}^{k} b_i L_i$$

where γ is the weight density of the material.

The weight of the frame is to be minimized subject to the following constraints:

1. Stress constraints at the ith node of Member 1:

$$\phi_i = \sigma_i - \sigma_{\max_1} \leq 0, \quad i = 1, ..., m \qquad (5\text{-}53)$$

where $\sigma_i = Mc(b_i)/[I(b_i)]$ is bending stress, $c(b_i) = \beta(b_i)^{1/2}$ is half the depth of the beam at point i, $i = 1, 2, \ldots, m$ and $\sigma_{\max_1}$ is the maximum allowable stress. The parameter β is a property of the cross-sectional geometry.

2. Deflection constraint:

$$\phi_{m+1} = u_A - \Delta < 0 \qquad (5\text{-}54)$$

where u_A is the horizontal deflection at the top of Member 1 and Δ is the maximum allowable lateral deflection of the top of the frame.

3. Buckling constraint:

$$\phi_{m+2} = \frac{3EI_3}{L^3} u_A - P < 0 \qquad (5\text{-}55)$$

where the first term is just the load R carried by the column.

4. Compressive-stress constraint at the jth node of Member 2:

$$\phi_{m+j+2} = \frac{P}{b_j} - \sigma_{max_2} < 0 \qquad (5\text{-}56)$$

for b_j in Member 2.

The steepest-descent algorithm can now be applied directly. Given data are: $\Delta = 4$ in.; $\sigma_{max_1} = 85{,}000$ psi; $\sigma_{max_2} = 5{,}000$ psi; cross-sectional area of Member 3 is 4 in.^{2}; $L = 100$ in.; $E = 3\times10^7$ psi, $\alpha = 0.07958$, and $\beta = 0.5642$. The resulting horizontal forces R that correspond to increasing constant lateral loads q, given in Table 5-7, are 285 lb, 409 lb, 458 lb, and 458 lb, respectively. For $q = 20$ and 25 lb/in., the displacement in Eq. 5-54 is an equality. For lower loads it is a strict inequality. This was determined automatically by the algorithm. The results for different side loadings are given in Tables 5-7 and 5-8. A profile of an optimal frame is shown in Fig. 5-12. Comput[illegible] in each case required approximat[illegible] per iteration and 15 iterations to converge on an IBM 360-65.

TABLE 5-7
OPTIMAL DESIGN VARIABLE b_i FOR STATIC FRAME

Cross-Sectional Area b_i, in.^{2}

Element	Member 1 q, lb/in.				Member 2 q, lb/in.			
No. i	10	15	20	25	10	15	20	25
1	1.49	2.03	2.76	4.51	0.226	0.271	0.286	0.286
2	0.86	1.20	1.76	3.34	0.366	0.438	0.464	0.464
3	0.40	0.47	0.48	2.10	0.407	0.487	0.516	0.516
4	0.43	0.53	0.37	0.43	0.366	0.438	0.464	0.464
5	0.43	0.53	0.52	0.43	0.226	0.271	0.286	0.286

TABLE 5-8
VOLUME OF OPTIMUM FRAME

	q, lb/in. 10	15	20	25
Optimum Volume, in.^{3}	504.2	533.4	558.0	656.0

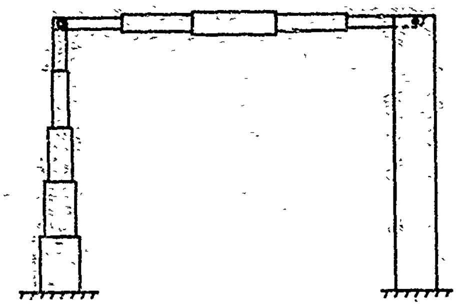


Figure 5-12. Profile of Optimal Frame With Multiple Failure Criteria ($q = 25$ lb/in.)

5-4.5 A MINIMUM WEIGHT PLATE WITH FREQUENCY CONSTRAINTS

As a final numerical example in this paragraph consider the problem of minimum weight design of the simply supported rectangular plate shown in Fig. 5-13 subject to a natural frequency constraint. The bending equation for plates of variable thickness is given in Eq. 5-58. When the deflection $W(x,y,t)$ is written in the form

$$W(x, y, t) = w(x, y) \cos \omega t \qquad (5\text{-}57)$$

the governing equation becomes

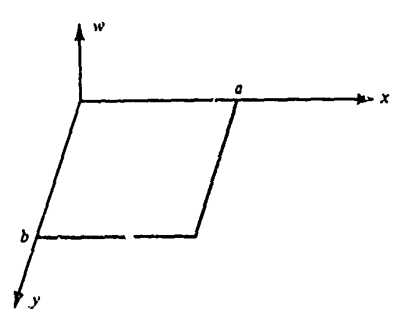


Figure 5-13. Rectangular Plate

$$D\nabla^4 w + 2\frac{\partial D}{\partial y}\frac{\partial}{\partial x}\nabla^2 w + 2\frac{\partial D}{\partial y}\frac{\partial}{\partial y}\nabla^2 w$$

$$+ \nabla^2 D \cdot \nabla^2 w - (1-\nu)\left[\frac{\partial^2 D}{\partial x^2}\frac{\partial^2 w}{\partial y^2} - 2\frac{\partial^2 D}{\partial x \partial y}\frac{\partial^2 w}{\partial x \partial y} + \frac{\partial^2 D}{\partial y^2}\frac{\partial^2 w}{\partial x^2}\right]$$

$$= h\rho\omega^2 w \qquad (5\text{-}58)$$

where

$$D(x,y) = \frac{Eh^3(x,y)}{12(1-\nu^2)} \qquad (5\text{-}59)$$

$h(x,y)$ is the thickness of the plate, which is the design variable, and ρ is the density of plate material.

When the function $w(x,y)$ is represented in the form

$$w(x,y) = \sum_{m,n=1} A_{mn} \sin\frac{m\pi x}{a} \sin\frac{n\pi y}{b} \qquad (5\text{-}60)$$

the eigenvalue problem can be solved approximately by numerical methods. The problem posed here is solved using a collocation technique, i.e., the differential equation is satisfied at discrete points in the region, Fig. 5-14.

The number of discrete points is chosen equal to the number of terms in the truncated series of Eq. 5-60. The drivatives of the function $D(x,y)$ at the grid points are evaluated by the use of finite differences. For a given set of design variables, i.e., $h(x,y)$, the lowest eigenvalue, $\zeta = \rho\omega^2$, and the associated eigenvector $\{A_{mn}\}$, which plays the role of y, are determined.

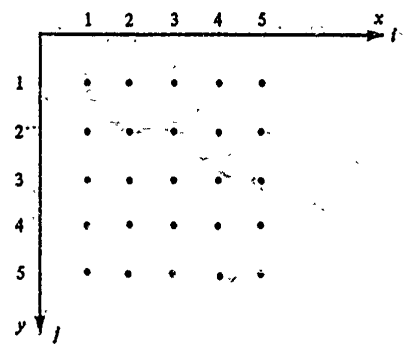


Figure 5-14. Collocation Points

In the steepest-descent algorithm, the cost function that is to be minimized is

$$J = \rho\Delta A \sum_{i,j=1} h(x_i,y_j) \qquad (5\text{-}61)$$

where ΔA is the area of the grid squares.

The constraints imposed on the design are

$$h_0 - h(x_i,y_j) < 0 \qquad (5\text{-}62)$$

and

$$\zeta_0 - \zeta < 0 \qquad (5\text{-}63)$$

where $h_0 > 0$ and $\zeta_0 > 0$ are lower limits on plate thickness and eigenvalue of Eq. 5-58, respectively.

The steepest-descent computational algorithm applies in a direct way. It should be noted that the collocation method for approximate solutions of the equations for natural frequency yields nonsymmetric matrices K and M in Eq. 5-16. Thus in this formulation of the plate optimization problem, the adjoint eigenvalue problem, Eq. 5-18, must be solved along with the original eigenvalue problem. If finite element methods for plate analysis had been used, symmetric matrices would have been obtained. In this example, as well as in the preceding, a minimum effort was expended to make computations efficient. The emphasis has been placed on getting results. A subsequent effort will be devoted to making algorithms more efficient.

The minimum weight plate problem was solved by the algorithm of par. 5-3 with $E = 3.0\times10^7$ psi, $\rho = 7.43\times10^{-4}$ lb-sec^2/in.^{4}, $\nu = 0.30$, and $\omega_0 = 1375$ rad/sec. The uniform plate with $\zeta = \rho\omega^2 = 1400$ was taken as the initial estimate to the optimization problem. The dimensions of the plate are 10.0 in. by 10.0 in. and the value of $h_0 = 0.1$ in. The material is assumed to have a constant density and so minimum weight is equivalent to minimum volume. The volume of the uniform plate is 11.44 in.^{3} and the volume of the optimal plate is 10.8 in.^{3} which is a 5.6% material savings. Fig. 5-15 shows 25 collocation points. The numbers in the network are

0 $\frac{a}{2}$ x

0.124	0.116	0.100	0.100	0.100
0.116	0.103	0.100	0.100	0.100
0.100	0.100	0.100	0.104	0.111
0.100	0.100	0.104	0.121	0.128
0.100	0.100	0.111	0.128	0.136

$\frac{b}{2}$ y

Figure 5-15. Optimal Design Variable h(x, y) for Vibrating Plate

the values of the thickness function $h(x,y)$ at each nodal point which is located at the center of each square. Double symmetry of the optimal plate thickness was observed about axes through the point $(a/2, b/2)$.

*This paragraph is based on the dissertation of Dr J. Arora, Ref. 34.

5-5 GENERAL TREATMENT OF TRUSS DESIGN*

The theory presented in pars. 5-2 and 5-3 will now be applied to the case of general plane and space trusses. These types of structures are encountered quite frequently in practical situations. Most common among these are buildings, transmission towers, bridges, cooling towers, aircraft structures, and lightweight military structures. In all these cases, it is desirable that the structure simultaneously should meet strength, deflection, and frequency requirements and be of minimum weight. In this chapter, all these constraints will be considered.

5-5.1 SPECIAL PROBLEM FORMULATION

In the problems to be considered here, geometry of the truss is assumed to be specified and the loads are applied only at the joints. The objective function for the problem is taken as the total weight or the volume of the truss, and the design variable for each member is taken as its cross-sectional area. The objective function of Eq. 5-1 in this case is a linear function of m design variables and may be written as

$$J = \sum_{i=1}^{m} \rho_i L_i b_i \tag{5-64}$$

where ρ_i and L_i are material density and length of member i, respectively.

The displacement method of structural analysis is used, and nodal displacements of the truss are considered as basic state variables. Therefore, the jth component of the state variable represents the jth displacement component of the truss. Fig. 5-16 shows a simple scheme of designating joints, members, and displacement components of a truss. Fig. 5-17 shows a bar element with sign conven-

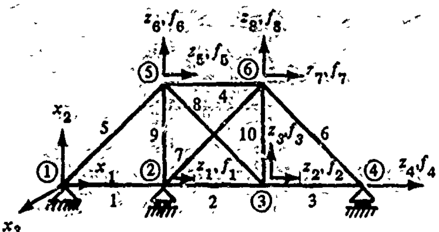


Figure 5-16. Description of a Truss

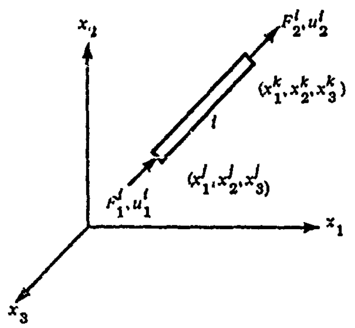


Figure 5-17. A Truss Element

tion to be used on element forces and deformations. Basic equations of the displacement method for a truss may be written as

$$u = \beta z \tag{5-65}$$

$$F = \bar{K}(b)u \tag{5-66}$$

and

$$f = \beta^T F \tag{5-67}$$

where u is the element deformation vector, F is the element force vector, f is the vector of external loads applied to structural nodes, and β is a rectangular transformation matrix, which transforms the nodal displacement vector z to the element deformation vector u. The matrix $\bar{K}(b)$ is composed of element stiffness matrices and is given by

$$\bar{K}(b) = \begin{bmatrix} \bar{K}_1 & & \\ & \ddots & \\ & & \bar{K}_m \end{bmatrix} \tag{5-68}$$

where m is the total number of elements in the truss and $\bar{K}_i$ is the stiffness matrix for the ith element of the truss. The stiffness matrix for the ith element may be written as

$$\bar{K}_i = \frac{E_i b_i}{L_i} \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \tag{5-69}$$

where E_i is Young's Modulus of Elasticity of the ith element. Substituting Eqs. 5-65 and 5-66 into Eq. 5-67, one obtains.

$$\begin{aligned} f &= [\beta^T \bar{K}(b)\beta]z \\ &= K(b)z \end{aligned} \tag{5-70}$$

where

$$K(b) = \beta^T \bar{K}(b)\beta \tag{5-71}$$

is the structure stiffness matrix, which is identical to $A(b)$ in Eq. 5-2. The mass matrix $M(b)$ for the truss may also be computed in a similar way, and it is given by

$$M(b) = \beta^T \bar{M}(b)\beta \tag{5-72}$$

where $\bar{M}(b)$ is formed from element mass matrices and is given by

$$\bar{M}(b) = \begin{bmatrix} \bar{M}_1 & & \\ & \ddots & \\ & & \bar{M}_m \end{bmatrix} \tag{5-73}$$

Here, $\bar{M}_i$ is the mass matrix for the ith element of the truss and is given by

$$\bar{M}_i = \frac{\rho_i b_i L_i}{6} \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}. \qquad (5\text{-}74)$$

Any nonstructural mass that is attached to the truss may also be added to the mass matrix of Eq. 5-72 and it may be written as

$$M(b) = \beta^T \bar{M}(b)\beta + M_o \qquad (5\text{-}75)$$

where M_o is a matrix consisting of nonstructural masses. In the example problems to be presented later. M_o is taken to be a null matrix. However, there is no particular difficulty in incorporating this matrix if it is not zero. Its inclusion will simply change the lowest natural frequency of the truss. The derivation of the given structural analysis equations and matrices is well documented in the literature (Refs. 20, 24).

In order to apply the algorithm of par. 5-3.2, two main matrices ℓ^J and ℓ^{ϕ} of Eqs. 5-20 and 5-21 must be computed. They can be assembled very easily once various other matrices required in them have been computed. In the class of problems treated here $f(b)$ does not depend on the design variable, so $\partial f(b)/\partial b = 0$. Also, one obtains from Eq. 5-64

$$\frac{\partial J}{\partial b} = (\rho_1 L_1, \ldots\ldots, \rho_m L_m)$$

$$\frac{\partial J}{\partial z} = (0, \ldots\ldots, 0)$$

and $\partial J/\partial \zeta = 0$; and from Eq. 5-14 $\lambda^J = (0, \ldots\ldots, 0)^T$. Substitution of these values into Eq. 5-20 yields

$$\ell^J = (\rho_1 L_1, \ldots\ldots, \rho_m L_m)^T \qquad (5\text{-}76)$$

Assembling the matrix ℓ^{ϕ} of Eq. 5-21 is more tedious. It requires a formation of constraint set ϕ as in Eq. 4-105 which will be discussed in detail now. The constraint set ϕ may be divided into five subsets, namely, frequency, stress, buckling and displacement constraints, and lower and upper bounds on the design variables. Explanation of these subsets follows one by one and, for each subset, matrices $\partial\phi/\partial\zeta$, $\partial\phi/\partial q$, and $\partial\phi/\partial b$ are computed.

5-5.1.1 FREQUENCY CONSTRAINTS

In the example problems, only one frequency constraint is considered. However, if other frequency constraints are also present, these may be treated in a similar way. Since the design variable vector b is available at any iteration, the matrices $K(b)$ and $M(b)$ are computed from Eqs. 5-71 and 5-72, respectively. The lowest eigenvalue ζ and the associated eigenvector y are then obtained from Eqs. 5-5 and 5-6, respectively. Premultiplying both sides of Eq. 5-6 by $K^{-1}(b)$, one obtains

$$K^{-1}(b)M(b)y = \frac{1}{\zeta}y = \gamma y \qquad (5\text{-}77)$$

where $\gamma = 1/\zeta$. The power method is used to find the largest eigenvalue γ of $K^{-1}(b)M(b)$. This method of obtaining the largest eigenvalue is quite efficient in the present problem, since a very good approximation to the eigenvector at each iteration, except for the first one, is available from the previous iteration. The lowest eigenvalue is then given by $\zeta = 1/\gamma$. The frequency constraint may now be written as

$$\zeta > \zeta_o \qquad (5\text{-}78)$$

where ζ_0 corresponds to a given frequency. In terms of the notation used in Eq.

5-4, Eq. 5-78 may be written as

$$\phi_s(b, z, \zeta) = \zeta_o - \zeta < 0 \qquad (5\text{-}79)$$

where s is a number assigned to this constraint. If this constraint is violated, then $\Delta\phi_s = -(\zeta_o - \zeta)$, $\partial\phi_s/\partial b = (0, \ldots\ldots, 0)$, $\partial\phi_s/\partial\zeta = -1$, and $\partial\phi_s/\partial z = (0, \ldots\ldots, 0)$.

5-5.1.2 STRESS CONSTRAINTS

Since the matrix $K(b)$ is available, Eq. 5-70 can be solved for the unknown displacement vector z. Element forces can then be computed from Eqs. 5-65 and 5-66. Substitution of Eq. 5-65 into Eq. 5-66 yields

$$F = Sz \qquad (5\text{-}80)$$

where $S = \bar{K}(b)\beta$. It may be noticed from Fig. 5-17 that two forces are specified for each element but the primary force remains constant throughout a bar element. Therefore, $F_2^i = -F_1^i$, where superscript i denotes the element number. Dimensions of the matrix S may be reduced from $2m \times n$ to $m \times n$ by using this relationship. Stresses in the members may now be calculated as

$$\sigma_i = \frac{F_i}{b_i}. \qquad (5\text{-}81)$$

Once the stress for each member becomes known, it is checked against the critical stress. A number of these stresses may be violated in a particular iteration. The stress constraint for the ith member may be written as

$$\sigma_i < \sigma_i^c \qquad (5\text{-}82)$$

where σ_i^c is the critical stress for member i. It should be noted that, in terms of the notations used in Fig. 5-17, compressive stress in a member is taken as positive and accordingly inequality (Eq. 5-82) holds in such cases. For tension members, σ_i and σ_i^c are negative; therefore, Eq. 5-82 is written as $\sigma_i^c < \sigma_i$. The expressions that follow are written for the case of compression members. For tension members similar expressions can be readily written. Inequality Eq. 5-82 may be written as

$$\phi_s(b, z, \zeta) = \sigma_i - \sigma_i^c < 0 \qquad (5\text{-}83)$$

where s is an index assigned to this constraint. In all subsequent constraint subsets, subscript s on ϕ will have the same meaning. If Eq. 5-83 is violated, then $\Delta\phi_s = -(\sigma_i - \sigma_i^c)$, $\partial\phi/\partial\zeta = 0$,

$$\frac{\partial\phi_s}{\partial b_j} = 0 \text{ for } i \neq j, \frac{\partial\phi_s}{\partial b_i} = \left(\frac{1}{b_i}\frac{\partial F_i}{\partial b_i} - \frac{1}{b_i^2}F_i\right),$$

$$\text{and } \frac{\partial\phi_s}{\partial z} = \left(\frac{\partial\sigma_i}{\partial z_1}, \ldots\ldots, \frac{\partial\sigma_i}{\partial z_n}\right)$$

where $\partial\sigma_i/\partial z_j$ and $\partial F_i/\partial b_i$ may be computed from Eq. 5-80.

5-5.1.3 BUCKLING CONSTRAINTS

Each compression member of the truss is also checked for the Euler buckling load given by

$$P_i^c = \frac{\pi^2 E_i I_i}{L_i^2} \qquad (5\text{-}84)$$

where P_i^c and I_i are the critical buckling load and moment of inertia of the ith member, respectively. It is assumed that the moment of inertia of the cross section of a member can be written as

$$I_i = \alpha_i b_i^2 \qquad (5\text{-}85)$$

where α_i is a constant depending upon the cross-sectional geometry of the ith member. This is a convenient way of expressing the moment of inertia in terms of the cross-sectional area of a member, because the constant α_i can be specified by the designer quite readily. Therefore, Eq. 5-84 may now be written as

$$P_i^c = \frac{\pi^2 E_i \alpha_i b_i^2}{L_i^2} = \theta_i b_i^2 \tag{5-86}$$

where $\theta_i = \pi^2 E_i \alpha_i / L_i^2$. Eq. 5-86 may be written in terms of the critical buckling stress σ_i^b as

$$\sigma_i^b = \theta_i b_i. \tag{5-87}$$

Now the buckling constraint for the ith compression member may be written as

$$\bar{\phi}_s(b, z, \zeta) = \sigma_i - \sigma_i^b < 0. \tag{5-88}$$

If this buckling constraint is violated, then $\Delta\bar{\phi}_s = -(\sigma_i - \sigma_i^b)$, $\partial\bar{\phi}_s / \partial\zeta = 0$,

$$\frac{\partial\bar{\phi}_s}{\partial b_j} = 0 \text{ for } i \neq j, \frac{\partial\bar{\phi}_s}{\partial b_i} = \left(\frac{1}{b_i}\frac{\partial F_i}{\partial b_i} - \frac{1}{b_i^2}F_i\right) - \theta_i,$$

$$\text{and } \frac{\partial\bar{\phi}_s}{\partial z} = \left(\frac{\partial\sigma_i}{\partial z_1}, \ldots\ldots, \frac{\partial\sigma_i}{\partial z_n}\right)$$

where $\partial F_i / \partial b_i$ and $\partial\sigma_i / \partial z_j$ may again be computed from Eq. 5-80. The buckling constraints on all other compression members are treated in a similar way.

5-5.1.4 DISPLACEMENT CONSTRAINTS

The displacement components are known at this stage; therefore, the constraints on them may be written as

$$|z_j| < z_j^a \tag{5-89}$$

where z_j^a is the maximum allowable jth component of displacement. If a particular component of displacement is positive, then Eq. 5-89 is written as $z_j - z_j^a < 0$; and if it is negative, then it is written as $z_j^a - z_j < 0$. The expressions that follow are written for the case of positive displacement, and similar expressions can also be written for the case of negative displacement. In terms of the notation of Eq. 5-4 the constraint for the positive displacement may be written as

$$\bar{\phi}_s(b, z, \zeta) = z_j - z_j^a < 0 \tag{5-90}$$

If the jth displacement component exceeds an allowable limit, then

$$\Delta\bar{\phi}_s = -(z_j - z_j^a), \frac{\partial\bar{\phi}_s}{\partial\zeta} = 0, \frac{\partial\bar{\phi}_s}{\partial b} = 0, \ldots\ldots, 0),$$

$$\text{and } \frac{\partial\bar{\phi}_s}{\partial z} = (0, \ldots\ldots, \underset{(j\text{th})}{1}, 0, \ldots\ 0).$$

All the displacement components are checked and any other violation is treated in a similar way.

5-5.1.5 BOUNDS ON DESIGN VARIABLES

It may be necessary to put upper and lower bounds on each design variable. This constraint may be demanded by many practical, architectural or structural considerations. Moreover, a lower limit on each design variable is required in the algorithm in order to avoid the attainment of unrealizable designs such as negative areas. This constraint may be written as

$$b_i^L < b_i < b_i^U \tag{5-91}$$

where b_i^L is the lower and b_i^U is the upper bound on the ith design variable. Inequality Eq. 5-91 may be split up into two parts:

(1) Lower Bound on Design Variables:

This constraint is written as $b_i^L < b_i$ or in terms of notation of Eq. 5-4

$$\bar{\phi}_s(b, z, \zeta) = b_i^L - b_i < 0 \tag{5-92}$$

Violation of this constraint yields,

$$\Delta\bar{\phi}_s = -(b_i^L - b_i), \ \frac{\partial\bar{\phi}_s}{\partial\zeta} = 0,$$

$$\frac{\partial\bar{\phi}_s}{\partial b} = (0, \ldots, 0, \underset{(i\text{th})}{-1}, 0, \ldots, 0), \text{ and}$$

$$\frac{\partial\bar{\phi}_s}{\partial z} = (0, \ldots\ldots, 0).$$

(2) Upper Bound on Design Variables:

This constraint is very similar to the previous one and in the notation of Eq. 5-4 it is written as

$$\bar{\phi}_s(b, z, \zeta) = b_i - b_i^U < 0 \tag{5-93}$$

If the upper bound on any design variable is violated, then

$$\Delta\bar{\phi}_s = -(b_i - b_i^U), \ \frac{\partial\bar{\phi}_s}{\partial\zeta} = 0,$$

$$\frac{\partial\bar{\phi}_s}{\partial b} = (0, \ldots, 0, \underset{(i\text{th})}{1}, 0, \ldots, 0), \text{ and}$$

$$\frac{\partial\bar{\phi}_s}{\partial z} = (0, \ldots, 0).$$

It may be noticed here that the cross-sectional area of any member of the truss may be assigned a predetermined value by putting the same upper and lower bound on it. This situation may be encountered in practice due to various reasons, and as shown in Example Problem par. 4-3.1, the present formulation handles it without any difficulty.

After all the constraints have been considered, the matrices $\partial\bar{\phi}/\partial b$, $\partial\bar{\phi}/\partial\zeta$, and $\partial\bar{\phi}/\partial z$ are available and $\lambda^{\bar{\phi}}$ can be solved from Eq. 5-15. This still does not allow the matrix $\ell^{\bar{\phi}}$ of Eq. 5-21 to be assembled. The following matrices must also be computed

$$\frac{\partial}{\partial b}[K(b)z] \tag{5-94}$$

$$\frac{\partial}{\partial b}[K(b)y] \tag{5-95}$$

and

$$\frac{\partial}{\partial b}[M(b)y] \,. \tag{5-96}$$

These matrices are assembled automatically from the quantities such as $K(b)$, $M(b)$, z, and y, which have previously been calculated in the computer. The procedure of computing the matrix of Eq. 5-94 will be explained here; the matrices of Eqs. 5-95 and 5-96 are calculated in an exactly similar manner. Eq. 5-71 may be written as (see Appendix B):

$$K(b) = \sum_{i=1}^{m} \beta^{i^T} \bar{K}_i \beta^i$$

where $\bar{K}_i$ is the only quantity which is a function of b. Now, Eq. 5-94 can be written as follows, by substituting the above expression for $K(b)$:

$$\frac{\partial}{\partial b}[K(b)z] = \frac{\partial}{\partial b}\left[\left(\sum_{i=1}^{m} \beta^{i^T} \bar{K}_i \beta^i\right) z\right]$$

$$= \sum_{i=1}^{m} \frac{\partial}{\partial b}\left[\left(\beta^{i^T} \bar{K}_i \beta^i\right) z\right] \tag{5-97}$$

It should be noted here that the summation sign in Eq. 5-97 represents the summation of m matrices of dimension $(n \times m)$. The quantity inside the differentiation sign is an n-dimensional vector whose components may be dependent upon the design variable vector b. Therefore, the quantity inside the summation sign is an $n \times m$ matrix for each index i. However, in the present case, since $\bar{K}_i$ is a function of only b_i, the computation of Eq. 5-97 is greatly simplified. Consideration of each i in Eq. 5-97 generates a $(n \times m)$ matrix whose only nonzero elements are in its ith column. Computation of Eq. 5-97 is performed quite readily and automatically in the computer.

All the information being available, the matrix $\ell^{\bar{\phi}}$ of Eq. 5-21 may now be assembled and the algorithm of par. 5-3.2 may be used to solve actual problems.

5-5.2 MULTIPLE LOADING CONDITIONS

Most structures are designed to withstand a multiple loading environment. This is quite reasonable, because only a certain combination of loads may act on the structure at a particular time. This situation is handled in the par. 5-5.1 formulation by expanding the state variable vector z to include all states. The element force vector f is also expanded accordingly. Formulation of displacement, stress, and buckling constraints must also take into consideration all states of the system. This is handled in the manner that follows. While formulating a particular displacement constraint, the value of displacement for each loading case is checked and each violation is entered into the reduced constraint vector $\bar{\phi}$. After this, the procedure of calculating the matrices $\partial\bar{\phi}/\partial b$, $\partial\bar{\phi}/\partial\zeta$, and $\partial\bar{\phi}/\partial z$ is the same as explained earlier. An exact same procedure is followed in treating stress and buckling constraints. This procedure of taking into consideration all the loading conditions has worked out quite satisfactorily in the example problems.

5-5.3 EXAMPLE PROBLEMS

Several trusses are designed by applying the procedure presented in this paragraph. A computer program, based on the algorithm stated previously, was written in FORTRAN IV. The computations were performed on the University of Iowa IBM 360-65 computer. The stiffness matrix for the structure was inverted by the Gauss-Jordan elimination procedure, and the power method was used to find the smallest eigenvalue.

Results for three typical trusses are presented here. All these structures were designed with stress, displacement, buckling, and frequency constraints. Examples 4-1 and 4-2, par. 4-1.1, are compared with results in Ref. 25. These were designed with and without frequency and buckling constraints in order to compare the results with Ref. 25. Example 4-3, par. 4-3.1, is treated in Ref. 26, and it was also designed with only stress constraints in one case to compare results with Ref. 26. All sample problems had lower limits on areas of the elements and Example 4-3 had upper limits. The program is general enough to handle different lower and upper bounds on stresses in an element, elements of different materials, and a different buckling parameter α_i for each element. The examples follow:

1. Example 5-1. Five-node Four-bar Truss

Fig. 5-18 shows the geometry and the dimensions of the truss. Input and output information is given in Table 5-9. In order to compare the results with those of Ref. 25, the

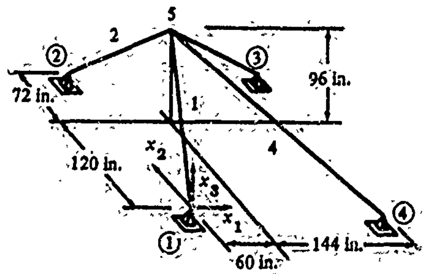


Figure 5-18. Four-bar Truss (Example 5-1)

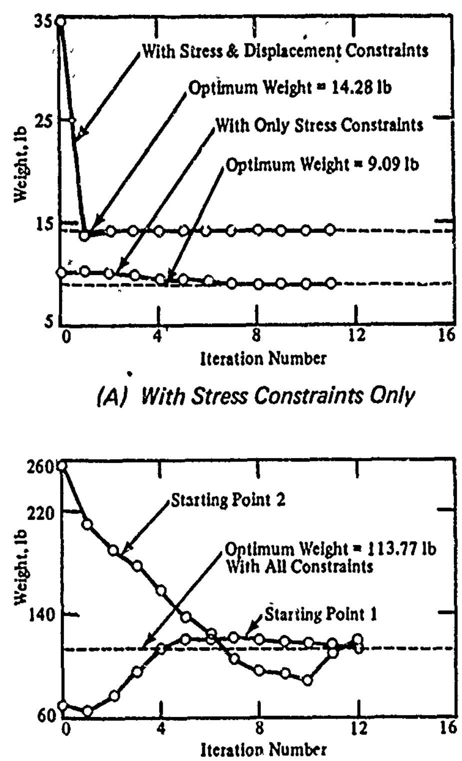
truss was first designed for stress constraints, Fig. 5-19(A), and second for stress and displacement constraints, Fig. 5-19(B). It may be noted that the results presented here are at least as good as those presented in Ref. 25.

The final design weight with only stress constraints was 9.09 lb and computation time was 1.820 sec for 12 iterations. The final design weight reported in Ref. 25 was 9.09 lb with a computation time of 4 sec for 5 cycles. The final design weight, with stress and displacement constraints, was 14.28 lb and the computation time was 1.500 sec for 12 iterations. The final design weight reported in Ref. 25 was 14.30 lb with a computation time of 10 sec for 4 cycles. It is difficult to make an exact comparison of the computation times because the computer used here is different from that used in Ref. 25. The computation times reported in Ref. 25 are on IBM 7094-II-7044 DCS Computer.

The truss was also designed by including buckling and frequency constraints along with other constraints. Two different starting points were used in optimizing this truss. Starting Point 1 was infeasible and Starting Point 2 was feasible. The final design weight beginning at Starting Point 1 was 113.48 lb and at Starting Point 2 was 113.77 lb. The slight difference in the two weights was due to the fact that in the first case the frequency constraint was violated by 0.154%, whereas in the second case this violation was 0.143%. Fig. 5-19 shows variation of the cost function with respect to the number of iterations for this problem. It may be noted that for practical purposes, convergence was obtained in six to eight iterations for all the cases.

2. Example 5-2: Transmission Tower

Fig. 5-20 shows the geometry and the dimensions of the tower. This example also is treated in Ref. 25. In this problem, the cross-sectional area of each member of the



(A) With Stress Constraints Only

(B) With All Constraints

Figure 5-19. Iteration vs Weight Curves for Example 5-1, Four-bar Truss

TABLE 5-9

FOUR-BAR TRUSS (EXAMPLE 5-1)

Design Information: For each member, Young's Modulus of Elasticity E_i, the specific weight ρ_i, lower limit on area of cross section b_i^L, and the constant α_i are 10^4 kips/in.^{2}, 0.10 lb/in.^{3}, 0.10 in.^{2}, and 1, respectively. There is no upper limit on member size. The resonant frequency for the truss is 284.6 Hz. For Output 1, the stress limits on each member are ±25.0 kips/in.^{2} and the displacement limits at node five are, 0.0, ±0.3 in., and ±0.4 in. in the x_1-, x_2-, and x_3-directions, respectively. For Output 2, the stress limits for each member are ±15.0 kips/in.^{2}, and the displacement limits at node five are 0.15 in. in all three directions. There are three loading conditions for the truss; they are: in positive x_1-, x_2-, and x_3- directions, 5, 0, 0; 0, 5, 0; and 0, 0, 7.5 kip, respectively, applied at node five.

OUTPUT 1. *With Stress and Displacement Constraints Only*

With only stress constraints Time per iteration = 0.152 sec Total time = 1.820 sec			With displacement constraints, also Time per iteration = 0.124 sec Total time = 1.500 sec		
El. No.	Starting Values, in.^{2}	Final Values, in.^{2}	El. No.	Starting Values, in.^{2}	Final Values, in.^{2}
1	0.100	0.130	1	0.500	0.234
2	0.200	0.192	2	0.500	0.319
3	0.200	0.120	3	0.500	0.184
4	0.100	0.100	4	0.500	0.128
Weight, lb	10.19	9.09	Weight, lb	34.86	14.28

OUTPUT 2. *With All Constraints*

Starting Point 1 Time per iteration = 0.147 sec Total time = 4.710 sec			Starting Point 2 Time per iteration = 0.172 sec Total time = 4.640 sec		
El. No.	Starting Values, in.^{2}	Final Values, in.^{2}	El. No.	Starting Values, in.^{2}	Final Values, in.^{2}
1	1.000	0.543	1	2.000	0.559
2	1.000	1.961	2	4.000	1.883
3	1.000	3.635	3	8.000	3.703
4	1.000	0.479	4	1.000	0.468
Weight, lb	69.72	113.48	Weight, lb	257.68	113.77

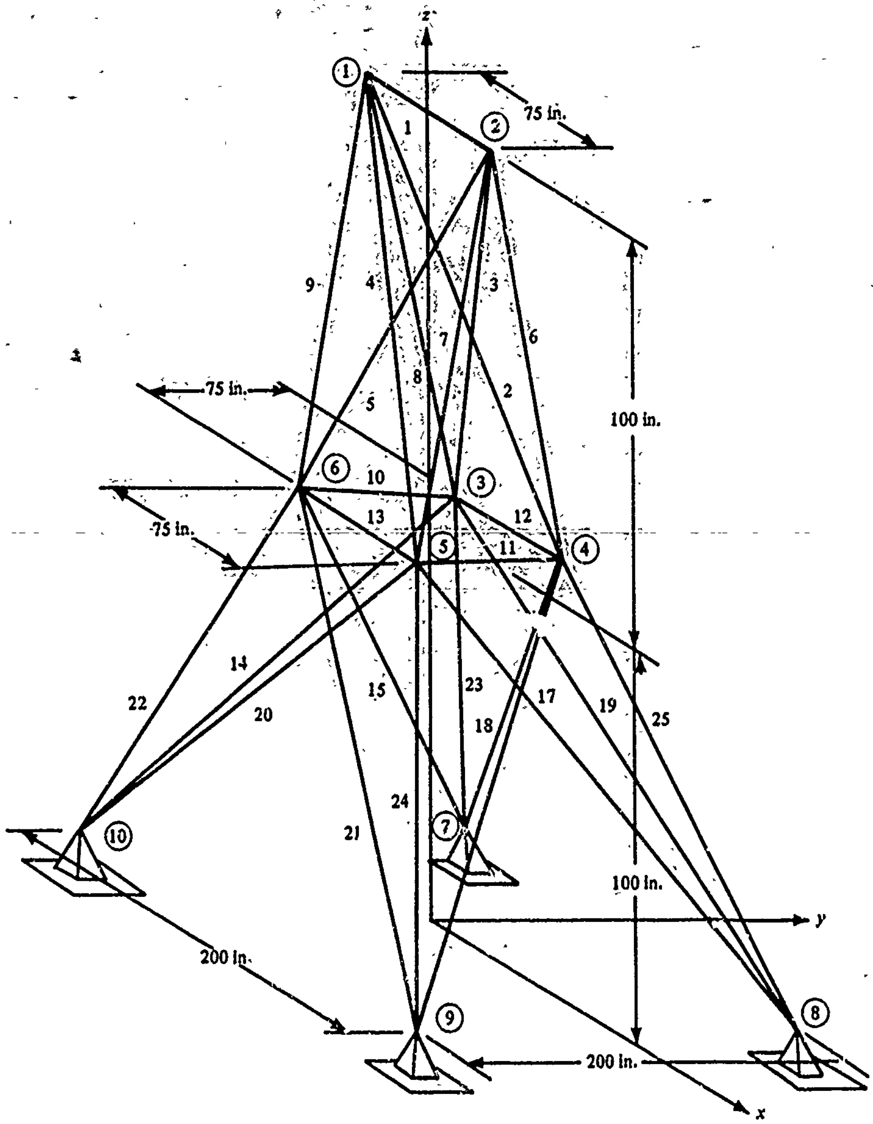


Figure 5-20. Transmission Tower (Example 5-2)

truss is treated as an unknown design variable, and the results obtained are given in Table 5-10. The tower was designed first with only stress constraints. The final design weight in this case was 91.13 lb with a computation time of 38 sec for 12 iterations. The final design weight reported in Ref. 25 was 91.14 lb with a computation time of 9 sec for 5 cycles. The values of final design variables compare quite well with those in Ref. 25. At the final design point all constraints were satisfied within 0.006%.

The tower was also designed with stress and displacement constraints and, finally, with all the constraints included. The design weight in

TABLE 5-10

TRANSMISSION TOWER (EXAMPLE 5-2)

Design Information: For each member of the structure, the modulus of elasticity E_i, the specific weight ρ_i, the constant α_i, and the stress limits are 10^4 kips/in.^{2}, 0.10 lb/in.^{3}, 1.0 and ± 40.0 kips/in.^{2}, respectively. The lower limit on the area of cross section of each member is 0.10 in.^{2} for the case with stress constraints only and 0.01 in.^{2} for other cases. There is no upper limit on the member sizes. The resonant frequency for the structure is 173.92 Hz, and the displacement limits are 0.35 in. on all nodes and in all directions. There are six loading conditions, and they are as follows (all loads are in kips):

Load Cond.	Node	Direction of Load			Load Cond.	Node	Direction of Load		
		x_1	x_2	x_3			x_1	x_2	x_3
1	1	1.0	10.0	−5.0	2	1	0	10.0	−5.0
	2	0	10.0	−5.0		2	−1.0	10.0	−5.0
	3	0.5	0	0		4	−0.5	0	0
	6	0.5	0	0		5	−0.5	0	0
3	1	1.0	−10.0	−5.0	4	1	0	−10.0	−5.0
	2	0	−10.0	−5.0		2	−1.0	−10.0	−5.0
	3	0.5	0	0		4	−0.5	0	0
	6	0.5	0	0		5	−0.5	0	0
5	1	0	20.0	−5.0	6	1	0	−20.0	−5.0
	2	0	−20.0	−5.0		2	0	20.0	−5.0

Output:

	With Stress Constraints Only		With Stress and Displacement Constraints		With All Constraints	
El. No.	Starting Values, in.^{2}	Final Values, in.^{2}	Starting Values, in.^{2}	Final Values, in.^{2}	Starting Values, in.^{2}	Final Values, in.^{2}
1	0.200	0.100	1.000	0.010	0.500	0.010
2	0.500	0.376	3.000	2.322	2.500	2.092
3	0.500	0.376	3.000	2.322	2.500	2.075
4	0.500	0.376	3.000	2.322	2.500	2.095
5	0.500	0.376	3.000	2.322	2.500	2.083
6	0.500	0.471	3.000	2.768	2.500	2.357
7	0.500	0.471	3.000	2.768	2.500	2.354
8	0.500	0.471	3.000	2.768	2.500	2.350
9	0.500	0.471	3.000	2.768	2.500	2.335
10	0.200	0.100	1.000	0.010	0.500	0.035
11	0.200	0.100	1.000	0.010	0.500	0.035
12	0.200	0.100	1.000	0.010	0.500	0.087
13	0.200	0.100	1.000	0.010	0.500	0.084
14	0.200	0.100	2.000	0.690	1.500	1.113
15	0.200	0.100	2.000	0.690	1.500	1.113
16	0.200	0.100	2.000	0.690	1.500	1.112
17	0.200	0.100	2.000	0.690	1.500	1.112

TABLE 5-10 (Cont'd.)

Output: (Cont'd.)

	With Stress Constraints Only		With Stress and Displacement Constraints		With All Constraints	
El. No.	Starting Values, in.^{2}	Final Values, in.^{2}	Starting Values, in.^{2}	Final Values, in.^{2}	Starting Values, in.^{2}	Final Values, in.^{2}
18	0.500	0.277	2.000	1.524	2.000	2.056
19	0.500	0.277	2.000	1.524	2.000	2.053
20	0.500	0.277	2.000	1.524	2.000	2.046
21	0.500	0.277	2.000	1.524	2.000	2.058
22	0.500	0.380	3.000	2.733	3.000	2.822
23	0.500	0.380	3.000	2.733	3.000	2.808
24	0.500	0.380	3.000	2.733	3.000	2.803
25	0.500	0.380	3.000	2.733	3.000	2.785
Weight, lb	132.37	91.13	772.24	546.18	669.80	590.32

the first case was 546.18 lb with a computation time of 47 sec for 17 iterations, and the maximum constraint violation was 0.00011%. The comparable design weight reported in Ref. 25 was 555.11 lb with a computation time of 24 sec for 7 cycles. This shows that, when displacement constraints are also included, the results obtained with the new gradient projection method are slightly better than those of Ref. 25. For a design with all the constraints included, the final weight was 590.32 lb with a computation time of 129 sec for 36 iterations, and the maximum violation of constraint was 0.028%. Fig. 5-21 shows variation of the cost function with respect to the iteration number for the last two cases of this problem. It may be noted that for practical purposes, convergence was obtained in only 6 iterations.

3. Example 5-3. 47-Bar Plane Truss

The schematic diagram of the structure with dimensions is shown in Fig. 5-22. This example is also treated in Ref. 26 where it is optimized for a single loading condition. The design information and the results are shown in Table 5-11. In order to compare the results with Ref. 26, the truss was first optimized with stress constraints only. The final design weight was 2,993.37 lb with a computation

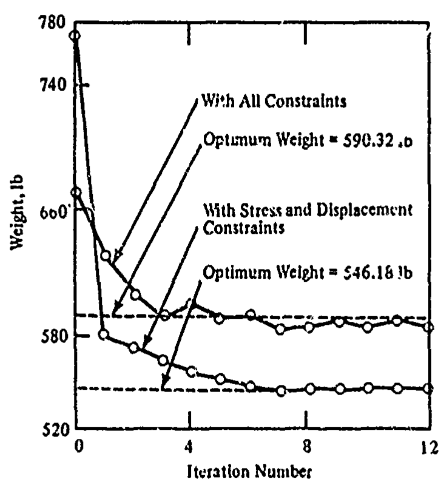


Figure 5-21. Iteration vs Weight Curves for Example 5-2, Transmission Tower

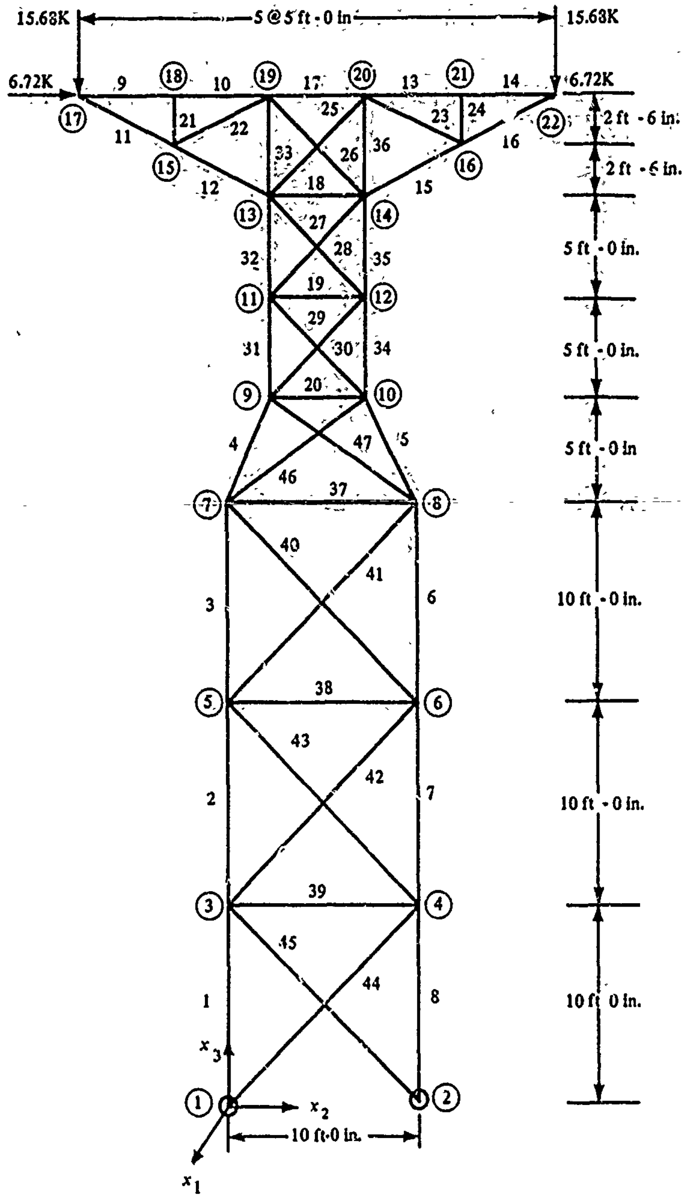


Figure 5-22. 47-Bar Plane Truss (Example 5-3)

time of 115 sec for 17 iterations. At this point the stress in member 18 was violated by 0.24% and all other violations were less than 0.035%. Another feasible design occurred at 9th iteration for which the design weight was 2,998.88 lb, maximum constraint violation was 0.10% for stress in 7th member and all other violations were less than 0.016%. The

TABLE 5-11

47-BAR PLANE TRUSS (EXAMPLE 5-3)

Design Information: For each member of the structure, the modulus of elasticity E_i, the specific weight ρ_i, and the constant α_i are 3.0×10^4 kips/in.^{2}, 0.284 lb/in.^{3}, and 1.0, respectively. The resonant frequency for the structure is 16.0 Hz and the displacement limits are 1 in. on all nodes and in all directions. There is one loading condition for the truss which is shown on Fig. 5-22. Allowable stress in tension for all members is 21.28 kips/in.^{2}

Output:

El. No.	Lower Area Bound, in.^{2}	Upper Area Bound, in.^{2}	Compression Stress Limit, kips/in.^{2}	Initial Area, in.^{2}	Final Area, in.^{2}	
					With Stress Constraints Only	With All Constraints
1	3.570	9.620	14.56	5.690	3.570	7.537
2	3.570	9.620	14.56	5.690	3.570	5.771
3	3.570	9.620	14.56	5.690	3.570	3.570
4	3.570	9.620	14.56	5.690	3.570	4.473
5	3.570	9.620	14.56	5.690	3.752	6.505
6	3.570	9.620	14.56	5.690	3.570	6.124
7	3.570	9.620	14.56	5.690	4.212	7.777
8	3.570	9.620	14.56	5.690	5.217	9.529
9	1.930	2.940	15.90	2.210	1.930	1.930
10	1.930	2.940	15.90	2.210	1.930	1.930
11	1.930	2.940	15.[illegible]0	2.210	2.205	2.199
12	1.930	2.940	15.90	2.210	2.205	2.205
13	1.930	2.940	15.90	2.210	1.930	2.940
14	1.930	2.940	15.[illegible]0	2.210	1.930	2.940
15	1.930	2.940	15.90	2.210	2.205	2.119
16	1.930	2.940	15.90	2.210	2.205	2.205
17	1.360	2.190	15.46	2.100	1.417	2.136
18	1.360	2.190	15.46	2.100	1.815	1.630
19	1.360	2.190	15.46	2.100	1.360	1.360
20	1.360	2.190	15.46	2.100	1.360	1.360
21	0.376	0.376	3.36	0.376	0.376	0.376
22	0.376	0.376	3.36	0.376	0.376	0.376
23	0.376	0.376	3.36	0.376	0.376	0.376
24	0.376	0.376	3.36	0.376	0.376	0.376
25	1.360	2.190	12.32	2.100	1.360	1.455
26	1.360	2.190	12.32	2.100	1.360	1.451
27	1.360	2.190	12.32	2.100	1.360	2.137
28	1.360	2.190	12.32	2.100	1.360	1.360
29	1.360	2.190	12.32	2.100	1.360	1.492
30	1.360	2.190	12.32	2.100	1.360	1.428
31	2.940	6.040	17.47	3.850	2.940	3.774
32	2.940	6.040	17.47	3.850	2.940	2.940
33	2.940	6.040	17.47	3.850	2.940	2.940
34	2.940	6.040	17.47	3.850	2.940	5.592
35	2.940	6.040	17.47	3.850	2.940	3.582
36	2.940	6.040	17.47	3.850	2.940	2.940
37	0.940	1.320	4.93	1.200	0.940	0.940

TABLE 5-11 (Cont'd.)

Output: (Cont'd.)

El. No.	Lower Area Bound, in.²	Upper Area Bound, in.²	Compression Stress Limit, kips/in.²	Initial Area, in.²	Final Area, in.² With Stress Constraints Only	Final Area, in.² With All Constraints
38	0.940	1.320	4.93	1.200	0.940	0.940
39	0.940	1.320	4.93	1.200	0.940	0.940
40	2.940	6.040	10.75	3.500	2.940	2.940
41	2.940	6.040	10.75	3.500	2.940	2.940
42	2.940	6.040	10.75	3.500	2.940	2.940
43	2.940	6.040	10.75	3.500	2.940	2.940
44	2.940	6.040	10.75	3.500	2.940	2.940
45	2.940	6.040	10.75	3.500	2.940	2.940
46	2.940	6.040	10.75	3.500	2.940	2.940
47	2.940	6.040	10.75	3.500	2.940	2.940
	Weight, lb			3910.30	2993.37	3771.0

final weight reported in Ref. 26 was 3,328.5 lb which is considerably higher than the one reported herein. This may be attributed to the fact that in Ref. 26 the members are divided into eight groups so that there are only eight independent design variables, whereas in this treatment, area of cross section of each member of the truss is treated as an unknown design variable.

The truss also was designed by imposing all the constraints. The starting point, stress limits, and upper and lower bounds on the areas are same as those used in Ref. 26. It may be noted that members 21, 22, 23, and 24 had the same upper and lower bounds on areas. The final design weight was 3,771.0 lb with a computation time of 166 sec for 24 iterations. The maximum violation of the constraint was 0.27% on stress for member 11. Fig. 5-23 shows variation of the cost function with respect to the number of iteration for both the cases. It may be noted that for practical purposes, convergence occurred in approximately 6 iterations.

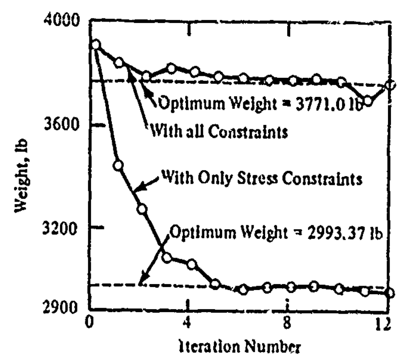


Figure 5-23. Iteration vs Weight Curves for Example 5-3, 47-Bar Plane Truss

5-6 A GENERAL TREATMENT OF PLANE FRAME DESIGN*

In this paragraph, an application of th gradient projection method to framed struc-

*This paragraph is based on the dissertation of Dr J Arora, Ref. 34.

tures, will be presented. Rigid frames are found quite frequently in practical situations, including building and vehicle structures. In the present work, optimization of planar framed structures using wide flange steel sections is considered under the assumption of elastic, working stress analysis. The AISC *Steel Construction Manual* (Ref. 27) is used for the properties of these sections. The constraints considered are stress, buckling, displacement, natural frequency, and restrictions on design variables.

5-6.1 PROBLEM FORMULATION

In the problems considered here, the geometry of the frame is assumed to be specified, i.e., lengths of the members or the joint coordinates are not treated as design variables. Multiple loading conditions for the structure are treated by the procedure explained in par. 5-4.2. The moment of inertia for each element is treated as the design variable; therefore, b is a vector whose ith component b_i is the moment of inertia of the ith element. In calculating weight or volume of the structure, element direct stresses, element bending stresses, area of cross section, and the section modulus of each element must be known. Also, in order to calculate the allowable compressive stress for an element, its least radius of gyration r_i must be known. These quantities are required as continuous functions, rather than discrete numbers, in the present formulation. Since the moment of inertia of each element is its only design variable, the quantities area of cross section, section modulus, and the least radius of gyration must be expressed in terms of the moment of inertia of the element. These relationships of the ith element are written as follows:

$$A_i = a_i b_i^{1/2} \tag{5-98}$$

$$Z_i = c_i b_i^{3/4} \tag{5-99}$$

and

$$r_i = d_i b_i^{1/4} \tag{5-100}$$

where A_i is the area of cross section, Z_i is the section modulus, r_i is the least radius of gyration of the ith element of rigid frame, and a_i, c_i, and d_i are constants. These constants can be found by plotting curves of area of cross section, section modulus, and the least radius of gyration versus the moment of inertia of various economical beam and column sections. These curves have been drawn by Nakamura (Ref. 28) for wide flange sections of AISC *Steel Construction Manual* (Ref. 27), and the same values are used in this handbook. This approach of obtaining continuous relationship for area of cross section, section modulus, least radius of gyration, and the moment of inertia has also been used by other researchers in their work (Refs. 29, 30, 31).

The objective function, Eq. 5-1, for this problem is again taken as the total weight of the frame which may be written as

$$J = \sum_{i=1}^{m} \rho_i L_i A_i = \sum_{i=1}^{m} \rho_i L_i a_i b_i^{1/2} \tag{5-101}$$

The displacement method of structural analysis is used, and nodal displacements of the frame are considered as basic state variables. Therefore, the jth component of the state variable represents the jth displacement component of the frame. Fig. 5-24 shows a simple scheme for designating joints, members, and displacement components of a frame in the structure coordinate system. Fig. 5-25 shows a frame element in the member coordinate system with the sign convention to be used on element forces and deformations. It may be

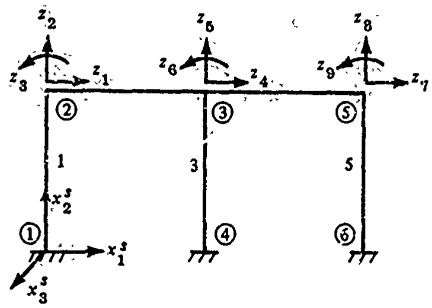


Figure 5-24. Description of a Frame

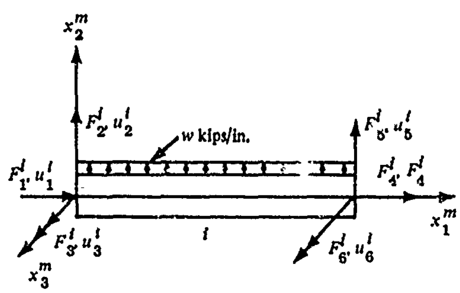


Figure 5-25. A Frame Element

noted that F_1^l and F_4^l are direct forces on the element, F_2^l and F_5^l are shearing forces, and F_3^l and F_6^l are the moments at the end of an element. The structural analysis equations developed in par. 5-4 and in Appendix B are also used here. The element forces are computed from Eq. 5-80 which may again be written as follows:

$$F = Sz \tag{5-102}$$

where

$$S = \bar{K}(b)\beta \tag{5-103}$$

Dimensions of the matrices $\bar{K}(b)$ and β are adjusted for the case of frames and F is a vector which consists of element forces for all the elements of the frame. The stiffness and mass matrices for the frame element are different from the truss element. They are given by the following matrices:

$$\bar{K}_l = \frac{E_l b_l}{L_l} \begin{bmatrix} a_l b_l^{-1/2} & 0 & 0 & -a_l b_l^{-1/2} & 0 & 0 \\ 0 & 12/L_l^2 & 6/L_l & 0 & -12/L_l^2 & 6/L_l \\ 0 & 6/L_l & 4 & 0 & -6/L_l & 2 \\ -a_l b_l^{-1/2} & 0 & 0 & a_l b_l^{-1/2} & 0 & 0 \\ 0 & -12/L_l^2 & -6/L_l & 0 & 12/L_l & 6/L_l \\ 0 & 6/L_l & 2 & 0 & -6/L_l & 4 \end{bmatrix} \tag{5-104}$$

$$\bar{M}_i = \frac{\rho_i L_i a_i b_i^{1/2}}{420} \begin{bmatrix} 140 & 0 & 0 & 70 & 0 & 0 \\ 0 & 156 & 22L & 0 & 54 & -13L \\ 0 & 22L & 4L^2 & 0 & 13L & -3L^2 \\ 70 & 0 & 0 & 140 & 0 & 0 \\ 0 & 54 & 13L & 0 & 156 & -22L \\ 0 & -13L & -3L^2 & 0 & -22L & 4L^2 \end{bmatrix} \quad (5\text{-}105)$$

Now, as before, the matrices ℓ^J and $\ell^{\bar{\phi}}$ of Eqs. 5-20 and 5-21, respectively, must be computed in order to apply the algorithm of par. 5-3.2. They can be readily assembled once various other matrices have been computed. Let us first consider computation of the matrix ℓ^J of Eq. 5-20. The matrix $f(b)$, which is computed from externally applied loads, is independent of the design variable vector b if the self weight of the elements is neglected. This implies that $\partial f(b)/\partial b = 0$. Also, from Eq. 5-101 one obtains

$$\frac{\partial J}{\partial b} = \frac{1}{2}\left(\rho_1 L_1 a_1 b_1^{-1/2}, \ldots\ldots, \rho_m L_m a_m b_m^{-1/2}\right) \quad (5\text{-}106)$$

$$\frac{\partial J}{\partial z} = (0, \ldots\ldots, 0), \quad (5\text{-}107)$$

and $\partial J/\partial \zeta = 0$. Eq. 5-14 now yields, $\lambda^J = (0, \ldots\ldots, 0)^T$. Substituting these values into Eq. 5-20, one obtains

$$\ell^J = \frac{1}{2}\left(\rho_1 L_1 a_1 b_1^{-1/2}, \ldots\ldots, \rho_m L_m a_m b_m^{-1/2}\right)^T \quad (5\text{-}108)$$

Next, consider computation of the matrix $\ell^{\bar{\phi}}$ of Eq 5-21. It requires formation of the constraint vector $\bar{\phi}$ by considering various constraints and computation of matrices such as $\partial\bar{\phi}/\partial\zeta$, $\partial\bar{\phi}/\partial z$, and $\partial\bar{\phi}/\partial b$. The treatment of frequency, displacement, and design variable constraints in the case of a frame is exactly the same as in the case of a truss, which is developed in par. 5-4. So, these features will not be explained here, except for the fact that any point where a displacement constraint must be imposed is treated as a nodal point. Computation of matrices such as $\partial/\partial b\ [K(b)z]$, $\partial/\partial b\ [K(b)y]$, and $\partial/\partial b\ [M(b)y]$ is also carried out in the way explained in par. 5-4. The only constraint that remains to be considered is the stress constraint, which will be considered next.

5-6.2 STRESS CONSTRAINT CALCULATIONS

Let s denote the subscript for this constraint. If one can compute $\partial\phi_s/\partial\zeta$ and row vectors $\partial\phi_s/\partial z$ and $\partial\phi_s/\partial b$ for this constraint, then he can assemble the matrix $\ell^{\bar{\phi}}$ of Eq. 5-21.

The members of a framed structure are subjected to direct as well as bending stresses. Thus, the effect of combined stresses must be considered in implementing the stress constraints. It should be noted here that a clear distinction is made between the elements and the members of a frame. This distinction is necessitated by the fact that a member must often be divided into several elements for structural analysis and implementation of displacement constraints. On the other hand, the compressive stress for all elements making up a member is the same. In the present work, the members subjected to direct and bending stresses are required to satisfy the AISC specification (Ref. 27). The permissible stress, according to this *Steel Construction Manual*, are:

1. Tension:

$$F_t = 0.60\,F_y \tag{5-109}$$

2. Bending:

$$F_b = 0.66\,F_y \tag{5-110}$$

where F_y is material yield stress, F_t is allowable tensile stress, and F_b is allowable bending stress.

3. Compression:

On the gross cross-sectional area of axially loaded compression members, when kL/r, the largest effective slenderness ratio of any unbraced segment, is less than C_c

$$F_a = \frac{\left(1 - \frac{1}{2}e^2\right) F_y}{\text{F.S.}} \tag{5-111}$$

where F.S. = factor of safety =

$$5/3 + (3/8)e - (1/8)e^3 \tag{5-112}$$

$$C_c = \sqrt{2\pi^2 E/F_y} \tag{5-113}$$

$$e = \frac{kL}{rC_c} \tag{5-114}$$

F_a = allowable compressive stress

E = Young's modulus

On the cross section of axially loaded columns when kL/r exceeds C_c,

$$F_a = \frac{1.49 \times 10^5}{(kL/r)^2}, \text{ Ksi (Kip/in}^2\text{)}. \tag{5-115}$$

4. Combined Stresses:

a. Axial Compression and Bending:

Members subjected to both axial compression and bending stresses shall be proportioned to satisfy the following requirements:

(1) When $f_a/F_a < 0.15$,

$$\frac{f_a}{F_a} + \frac{f_b}{F_b} < 1.0 \tag{5-116}$$

(2) When $f_a/F_a > 0.15$,

$$\frac{f_a}{F_a} + \frac{C_m f_b}{\left(1 - \frac{f_a}{F_e'}\right) F_b} < 1.0 \tag{5-117}$$

and, in addition at points braced in the plane of bending,

$$\frac{f_a}{0.6F_y} + \frac{f_b}{F_b} < 1.0 \tag{5-118}$$

where

F_a = axial stress that would be permitted if axial force alone existed

F_b = compressive bending stress that would be permitted if bending moment alone existed

$$F'_e = \frac{1.49 \times 10^5}{(kL_b/r_b)^2}, \text{Ksi} \qquad (5\text{-}119)$$

(In Eq. 5-119, for F'_e, L_b is the actual unbraced length (in.) in the plane of bending, r_b is the corresponding radius of gyration (in.), and k is the effective length factor in the plane of bending.)

f_a = computed axial stress

f_b = computed compressive bending stress at the point under consideration

C_m = a coefficient whose value shall be taken as follows:

(a) For compression members in frames subject to joint translation (sidesway):

$$C_m = 0.85. \qquad (5\text{-}120)$$

(b) For restrained compression members in frames braced against joint translation and not subject to transverse loading between their supports in the plane of bending:

$$C_m = 0.6 + 0.4\frac{M_1}{M_2}$$

but not less than 0.4, (5-121)

where M_1/M_2 is the ratio of the smaller to larger moments at the ends of that portion of the member, unbraced in the plane of bending under consideration. M_1/M_2 is positive when the member is bent in single curvature and negative when it is bent in reverse curvature.

(c) For compression members in frames braced against joint translation in the plane of loading and subjected to transverse loading between their supports, the value of C_m may be determined by rational analysis. However, in lieu of such analysis, the following values may be used.

1. For members whose ends are restrained:

$$C_m = 0.85$$

2. For members whose ends are unrestrained:

$$C_m = 1.0 \qquad (5\text{-}122)$$

b. Axial Tension and Bending

Members subject to both axial tension and bending stresses shall be proportioned to satisfy the requirements of Eq. 5-118 where f_b and F_b are taken, respectively, as the computed and permitted bending tensile stress.

Eqs. 5-116, 5-117 and 5-118 are known as the interaction equations. These equations, of course, are derived from the linear superposition of the direct stress under axial load alone and the bending stress under bending moment alone. The factor $C_m/(1 - f_a/F'_e)$ is used in Eq. 5-117 to account for the magnification of the primary bending moment due to the axial load. This factor depends upon the type of loading and end conditions of the member. The value of the coefficient C_m can be derived for various types of loadings and members, but the values recommended in Eqs. 5-120, 5-121, and 5-122 are conservative and are used in the present work. For a detailed development and discussion of these equations the reader is referred to Ref. 32.

The allowable compressive stress formula, Eq. 5-115, is derived based on the basic theory of column buckling. It is obtained by dividing the Euler buckling stress by a factor of safety of 1.92. Therefore, $F_a = \pi^2 E/[1.92 \times (kL/r)^2]$ and, taking $E = 3.0 \times 10^4$ Ksi, $F_a = 1.49 \times 10^5/(kL/r)^2$. Eq. 5-115 is applicable when the largest slenderness ratio kL/r is greater than or equal to C_c. Experiments have shown that when $kL/r < C_c$, the values of the failure stress predicted by the Euler critical stress formula are seldom attained (Ref. 32). This is due to the presence of residual stresses and other imperfections in fabrication of the members. Therefore, when $kL/r < C_c$, the values of the allowable stress F_a are found from Eq. 5-111 which is derived based on the parabolic approximation of the curve–critical stress F_a versus the slenderness ratio kL/r in the range $kL/r < C_c$. This approximation is chosen based on the experimental results obtained at Lehigh University (Ref. 32). The value of the constant C_c is found by assuming that the Euler critical stress formula holds until the critical stress is $F_y/2$. Therefore,

$$F_y/2 = \frac{\pi^2 E}{(kL/r)_c^2} \quad \text{or } C_c = (kL/r)_c^2 = \sqrt{2\pi^2 E/F_y}$$

where L and r must be expressed in the same units.

The factor of safety is used to account for small imperfections of form and loading, and variations of support and restraint conditions from those assumed in computation, which cause the true effective length to be different from that calculated. The factor of safety given by Eq. 5-112 includes an allowance for both of these factors and is adjusted to account for their varying influence. For short columns, Eq. 5-112 approaches the basic safety factor in tension (1.67); and, at $e = 1$, it becomes 15% higher (1.92), a value which is then used in the case when kL/r exceeds C_c. Eq. 5-112 is an approximation of a quarter sine wave between the two limits, the curve used in the specification as best representing the influence of the two factors. For a detailed discussion of the these factors, the reader is again referred to Ref. 32.

The effective length factor k for each member of the frame is found from the differential equation

$$\frac{d^2y}{dx^2} + \frac{Py}{EI} = 0, \tag{5-123}$$

where P is the buckling load, and I is the second moment of the cross-sectional area.

The solution of this equation is given by

$$y(x) = D_1 \sin\sqrt{\frac{P}{EI}}\, x + D_2 \cos\sqrt{\frac{P}{EI}}\, x \tag{5-129}$$

In rigid frames, two cases must be discussed: (1) frames without sidesway, and (2) frames with sidesway. The transcedental equation that comes from Eq. 5-124, while satisfying the boundary conditions for a member of the frame without sidesway, is given by Ref. 32

$$\left[\frac{1}{2}(G_A + G_B + \frac{1}{4} G_A G_B (\pi/k)^2 - 1\right](\pi/k)$$

$$\times \sin(\pi/k) - \left[\frac{1}{2}(G_A + G_B)(\pi/k)^2 + 2\right]$$

$$\times \cos(\pi/k) + 2 = 0 \tag{5-125}$$

where G_A and G_B are given by the following equations:

$$G_A = \frac{\Sigma I_{cA}/L_{cA}}{\Sigma I_{bA}/L_{bA}} \tag{5-126}$$

and

$$G_B = \frac{\Sigma I_{cB}/L_{cB}}{\Sigma I_{bB}/L_{bB}} \tag{5-127}$$

The subscripts A and B refer to the two ends of the member under consideration and the subscripts c and b refer to the compressed and restraint members, respectively. For G_A, the summations extend over all members that are connected to joint A and for G_B the summations extend over all the members that are connected to joint B. So, for the first case of a frame without sidesway, the value of the effective length factor k must be found by solving the transcedental Eq. 5-125 for each member of the frame.

For the second case, i.e., a rigid frame with sidesway, the transcedental equation that comes out of Eq. 5-124 – while satisfying the boundary condition for a member of the frame–is given by

$$[G_A G_B (\pi/k)^2 - 36] \sin(\pi/k) - 6(G_A + G_B)(\pi/k)\cos(\pi/k) = 0, \tag{5-128}$$

where G_A and G_B are given by Eqs. 5-126 and 5-127, respectively. Thus, for the case of a frame with sidesway, Eq. 5-128 must be solved for k for each member of the frame. The secant method of nonlinear algebraic equations is used in finding the roots of Eqs. 5-125 and 5-128 in the present work.

The interaction Eqs. 5-116, 5-117, and 5-118 are implemented at the point of maximum bending moment for an element. If there are no loads between the end points of an element, then the maximum bending moment is at one of the ends; otherwise the actual point of maximum bending moment is found and the interaction equations are implemented there. As an example, consider the case of a uniformly distributed load on a frame element (Fig. 5-25); the moment at a distance x from the left end is given by

$$M_x = F_3^l - F_2^l x - \frac{wx^2}{2} \tag{5-129}$$

and

$$\frac{\partial M_x}{\partial x} = -F_2^l - wx = 0 \tag{5-130}$$

or

$$x_{max} = -\frac{F_2^l}{w} \tag{5-131}$$

Therefore, from Eq. 5-129

$$M_{max} = -F_3^l + (F_2^l)^2/(2w)\,. \tag{5-132}$$

Eq. 5-132 is used in computing the maximum bending stress required in the interaction equations. Now, the implementation of the interaction equations will be considered one by one and the vectors $\partial\bar{\phi}_s/\partial z$ and $\partial\bar{\phi}_s/\partial b$ will be computed in each case. For the sake of simplicity, let N be the direct force on the element, M_{max} be the maximum bending moment, k_l be the effective length factor, and L_l be the length of the member to which the lth element belongs.

b. Interaction Equations:

(1) Interaction Eq. 5-116:

Eq. 5-116, for the maximum stress in the ith element, can be written as

$$\bar{\phi}_s = \frac{N}{A_i F_a} + \frac{M_{max}}{Z_i F_b} - 1.0 < 0. \qquad (5\text{-}133)$$

where Z_i is the beam bending stiffness.

In case Eq. 5-133 is violated, then one must compute $\Delta\bar{\phi}_s$, $\partial\bar{\phi}_s/\partial\xi$, $\partial\bar{\phi}_s/\partial z$, and $\partial\bar{\phi}_s/\partial b$. Therefore, from Eq. 5-133 $\partial\bar{\phi}_s/\partial z = 0$ and

$$\Delta\bar{\phi}_s = -\left(\frac{N}{A_i F_a} + \frac{M_{max}}{Z_i F_b} - 1.0\right) < 0. \qquad (5\text{-}134)$$

Differentiating Eq. 5-133 with respect to z,

$$\partial\dot{\bar{\phi}}_s = \frac{1}{A_i F_a}\frac{\partial N}{\partial z} + \frac{1}{Z_i F_b}\frac{\partial M_{max}}{\partial z} \qquad (5\text{-}135)$$

The value of $\partial M_{max}/\partial z$ depends on the expression that defines M_{max}. If M_{max} occurs at an end of the element, then

$$M_{max} = F_3^i \text{ or } F_6^i \qquad (5\text{-}136)$$

and

$$\frac{\partial M_{max}}{\partial z} = \frac{\partial F_3^i}{\partial z} \text{ or } \frac{\partial F_6^i}{\partial z} \qquad (5\text{-}137)$$

If M_{max} occurs at a point other than the ends, then Eq. 5-132 gives its value, and

$$\frac{\partial M_{max}}{\partial z} = \frac{\partial F_3^i}{\partial z} + \frac{F_2^i}{w}\frac{\partial F_2^i}{\partial z}. \qquad (5\text{-}138)$$

Again, differentiating Eq. 5-133 with respect to b,

$$\frac{\partial\bar{\phi}_s}{\partial b} = \frac{1}{A_i F_a}\left(\frac{\partial N}{\partial b} - \frac{N}{2b_i}\frac{\partial b_i}{\partial b}\right) - \frac{N}{A_i F_a^2}\frac{\partial F_a}{\partial b} + \frac{1}{Z_i F_b} \times \left(\frac{\partial M_{max}}{\partial b} - \frac{3M_{max}}{4b_i}\frac{\partial b_i}{\partial b}\right) \qquad (5\text{-}139)$$

The value of $\partial M_{max}/\partial b$ can be found from Eq. 5-136 or Eq. 5-132, which are as follows:

$$\frac{\partial M_{max}}{\partial b} = \frac{\partial F_3^i}{\partial b} \text{ or } \frac{\partial F_6^i}{\partial b} \qquad (5\text{-}140)$$

$$\frac{\partial M_{max}}{\partial b} = \frac{\partial F_3^i}{\partial b} + \frac{F_2^i}{w}\frac{\partial F_2^i}{\partial b} \qquad (5\text{-}141)$$

The value of $\partial F_a/\partial b$ required in Eq. 5-139 is found from Eq. 5-111 or Eq. 5-115. First, if $k_i L_i/r_i < C_c$, then Eq. 5-111 gives the value of F_a and

$$\frac{\partial F_a}{\partial b} = \frac{F_y e_i}{4b_i (\text{F.S.})}\left[e_i + \frac{3}{8(\text{F.S.})} \times (1 - e_i^2)\left(1 - \frac{1}{2}e_i^2\right)\right]\frac{\partial b_i}{\partial b} - \frac{F_y e_i}{k_i (\text{F.S.})}\left[e_i + \frac{3}{8(\text{F.S.})} \times (1 - e_i^2)\left(1 - \frac{1}{2}e_i^2\right)\frac{\partial k_i}{\partial b}\right] \qquad (5\text{-}142)$$

where

$$e_i = \frac{k_i L_i}{r_i C_c} \qquad (5\text{-}143)$$

When $k_l L_l / r_l > C_c$, Eq. 5-115 gives the value of F_a and

$$\frac{\partial F_a}{\partial b} = \frac{F_a}{2b_l}\frac{\partial b_l}{\partial b} - \frac{2F_a}{k_l}\frac{\partial k_l}{\partial b}. \quad (5\text{-}144)$$

Substituting the appropriate expressions in Eq. 5-139, the value of $\partial\phi_s/\partial b$ can be found.

(2) Interaction Eq. 5-117:

Eq. 5-117 for the maximum stress in the lth element can be written as follows:

$$\bar{\phi}_s = \frac{N}{A_l F_a} + \frac{C_m M_{max}}{\psi_1} - 1.0 < 0 \quad (5\text{-}145)$$

where

$$\psi_1 = \left(1 - \frac{N}{A_l F_e'}\right) Z_l F_b. \quad (5\text{-}146)$$

If this constraint is violated, then one must compute $\Delta\bar{\phi}_s$, $\partial\bar{\phi}_s/\partial\zeta$, $\partial\bar{\phi}_s/\partial z$, and $\partial\bar{\phi}_s/\partial b$. In this case $\partial\bar{\phi}_s/\partial\zeta = 0$ and

$$\Delta\bar{\phi}_s = -\left(\frac{N}{A_l F_a} + \frac{C_m M_{max}}{\psi_1} - 1.0\right). \quad (5\text{-}147)$$

Differentiating Eq. 5-145 with respect to z and b, one obtains

$$\frac{\partial\bar{\phi}_s}{\partial z} = \left(\frac{1}{A_l F_a} + \frac{C_m M_{max}}{\psi_1 \psi_2 A_l F_e'}\right)\frac{\partial N}{\partial z} + \left(C_m \frac{\partial M_{max}}{\partial z} + M_{max}\frac{\partial C_m}{\partial z}\right)\Big/\psi_1 \quad (5\text{-}148)$$

and

$$\frac{\partial\bar{\phi}_s}{\partial b} = \frac{1}{A_l F_a}\left(\frac{\partial N}{\partial b} - \frac{N}{2b_l}\frac{\partial b_l}{\partial b}\right) - \frac{N}{A_l F_a^2}\frac{\partial F_a}{\partial b} + \left(C_m \frac{\partial M_{max}}{\partial b} + M_{max}\frac{\partial C_m}{\partial b}\right)\Big/\psi_1 - \frac{C_m M_{max}}{\psi_1 \psi_2}\left[\frac{3}{4b_l}(1 + N/A_l F_e')\frac{\partial b_l}{\partial b} - \frac{1}{A_l F_e'}\frac{\partial N}{\partial b} - \frac{2N}{k_l A_l F_e'}\frac{\partial k_l}{\partial b}\right] \quad (5\text{-}149)$$

where

$$\psi_2 = \left(1 - \frac{N}{A_l F_e'}\right). \quad (5\text{-}150)$$

(3) Interaction Eq. 5-118:

Next, consider Eq. 5-118, which may be written as follows for the maximum stress in the lth element

$$\bar{\phi}_s = \frac{N}{0.6F_y A_l} + \frac{M_{max}}{Z_l F_b} - 1.0 < 0. \quad (5\text{-}151)$$

Therefore, $\partial\bar{\phi}_s/\partial\zeta = 0$ and

$$\Delta\bar{\phi}_s = -\left(\frac{N}{0.6F_y A_l} + \frac{M_{max}}{Z_l F_b} - 1.0\right). \quad (5\text{-}152)$$

Differentiating Eq. 5-151 with respect to z and b, one obtains

$$\frac{\partial\bar{\phi}_s}{\partial z} = \frac{1}{0.6F_y A_l}\frac{\partial N}{\partial z} + \frac{1}{Z_l F_b}\frac{\partial M_{max}}{\partial z} \quad (5\text{-}153)$$

and

$$\frac{\partial\bar{\phi}_s}{\partial b} = \frac{1}{0.6F_y A_l}\left(\frac{\partial N}{\partial b} - \frac{N}{2b_l}\frac{\partial b_l}{\partial b}\right) + \frac{1}{Z_l F_b}\left(\frac{\partial M_{max}}{\partial b} - \frac{3M_{max}}{4b_l}\frac{\partial b_l}{\partial b}\right). \quad (5\text{-}154)$$

It may be noted in the previous equations that

$$\frac{\partial b_i}{\partial b} = (0, ..., 0, \overset{(i\text{th})}{1}, 0, ..., 0) \qquad (5\text{-}155)$$

The value of the vectors such as $\partial N/\partial z$, $\partial N/\partial b$, $\partial F_2^i/\partial z$, $\partial F_2^i/\partial b$ can be found directly from the Eq. 5-102. The vectors $\partial C_m/\partial z$ and $\partial C_m^-/\partial b$ are zero for cases prescribed by Eqs. 5-120 and 5-122. For the case prescribed by Eq. 5-121, they are computed using the chain rule of differentiation. It remains to find value of the vector $\partial k_i/\partial b$. This vector can be computed by differentiating Eq. 5-125 or Eq. 5-128 with respect to the design variable vector b. However, due to the fact that both G_A and G_B are functions of b, this computation is quite tedious and time consuming on the computer. Another approach that may be followed for computing $\partial k_i/\partial b$ is to use the method of finite differences; but this approach is equally time consuming on the computer. Moreover, it has been observed in the numerical computation that the value of k_i does not change appreciably from one design cycle to another. Therefore, without significant loss of accuracy, the value of k_i in a particular design cycle is treated as a constant. However, at the start of each design cycle, k values for all the members of the frame are recomputed. Thus, following this procedure, $\partial k_i/\partial b = (0, ..., 0)$. Now, all the necessary information is available to assemble matrix Λ of Eq. 5-20.

5-6.3 EXAMPLE PROBLEMS

Several rigid frames were optimized using the computer program based on the algorithm of par. 5-3.1. All the problems were solved with stress, displacement, frequency, and design variable constraints. Example problems 5-4 and 5-5 that follow also are treated in Ref. 28 and were first designed for only stress constraints in order to compare results with those of Ref. 28.

1. *Example 5-4. Simple Portal Frame*

Fig. 5-26 shows the dimension of the frame. The moment of inertia for each element of the frame is treated as an unknown, and the results obtained are shown in Table 5-12. The frame was first designed with only stress constraints. The final weight in this case was 3050.5 lb with a computation time of 3.74 sec for 13 cycles. At the final design point, the maximum constraint violation was 0.012% for stress in element 2. Optimal weight reported in Ref. 28 was 3206 lb, which is higher by approximately 5%.

The frame was also designed by including all the constraints. The resonant frequency limit for the structure was 25.0 Hz and the final weight obtained in this case was 3803.0 lb with a computation time of 14.60 sec for 31 iterations. At the final design point, the maximum constraint violation was 0.0073%

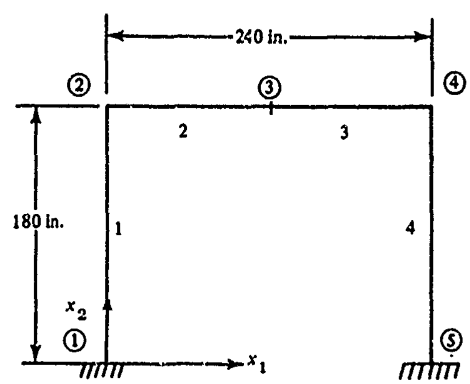


Figure 5-26. Simple Portal Frame (Example 5-4)

TABLE 5-12

SIMPLE PORTAL FRAME (EXAMPLE 5-4)

Design Information: For each element of the frame, the modulus of elasticity, the specific weight, and the yield stress are 3×10^4 kips/in.^{2}, 0.2836 lb/in.^{3}, and 36.0 kips/in.^{2}, respectively. The constants a_i, c_i, and d_i are 0.58, 0.58, and 0.67, respectively. The lower limit on the moment of inertia of each element is 1.0 in.^{4} and there is no upper limit. The resonant frequency is 25.0 Hz and the displacement limits are 0.5 in. at nodes 2, 3, and 4 in both x_1- and x_2-directions. There are three loading conditions for the frame; first is uniformly distributed load of -0.5 kip/in. on elements 2 and 3, second is a load of 45.0 kips in x-direction at node 2, and the third is a load of -45.0 kips in x_1-direction at node 4.

With Only Stress Constraints Computation time = 3.74 sec			With All the Constraints Computation time = 14.60 sec		
El. No.	Starting Values, in.^{4}	Final Values, in.^{4}	El. No.	Starting Values, in.^{4}	Final Values, in.^{4}
1	1600.0	1091.4	1	1600.0	1995.5
2	1600.0	768.3	2	1600.0	860.3
3	1600.0	768.3	3	1600.0	860.3
4	1600.0	1091.3	4	1600.0	1995.5
Wt, lb	3947.7	3050.5	Wt, lb	3947.7	3803.0

for stress in element 2. Fig. 5-27 shows the variation of the objective function as the iterations progress. It may be noted that, for practical purposes, convergence was obtained in only 5 iterations in the first case and in 7 iterations in the second case. However, in the second case, the cost function continued to reduce for a few cycles beyond the 7th iteration without correcting the constraints. This was due to the fact that the step size for the problem was too large.

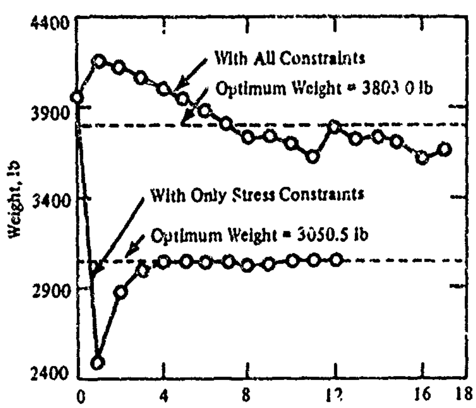


Figure 5-27. Iteration vs Weight Curves for Example 5-4, Simple Portal Frame

2. *Example 5-5. One-bay Two-story Frame*

The present example is also treated in Ref. 28. Fig. 5-28 shows the dimensions and the loading conditions for this structure. Input and output information for this example is given in Table 5-13. This frame was first designed for stress constraints only. The final weight in this case was 8292.0 lb with a computation time of 21.47 sec for 32 iterations. Maximum constraint violation at the design point in this case was 0.27 percent for stress in element 3. The comparable final weight reported in Ref. 28 was 8810 lb, which is again higher by approximately 5.8%.

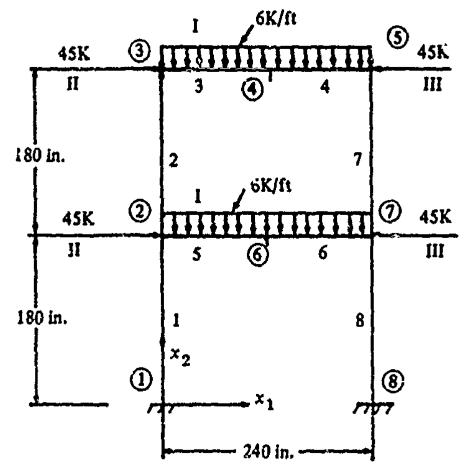


Figure 5-28. One-bay, Two-story Frame (Example 5-5)

The frame was also designed while enforcing all constraints. The resonant frequency limit in this case was 15.0 Hz. The final weight obtained in this case was 9722.5 lb with a computation time of 48.84 sec for 32 iterations. Maximum constraint violation was 0.38×10^{-3} % for displacement of node 3 in the x_1-direction. Fig. 5-29 shows variation of the cost function with respect to iteration number, and it may again be noted that convergence was obtained in 8 cycles in both the cases.

3. *Example 5-6. Two-bay Six-story Frame*

Figure 5-30 shows the geometry and dimensions of the frame. This frame has 21 joints, 30 members, and 54 degrees of freedom. The frame was designed for four loading conditions, and the input and output informa-

TABLE 5-13

ONE-BAY, TWO-STORY FRAME (EXAMPLE 5-5)

Design Information: For each element of the frame, the modulus of elasticity, the specific weight, and the yield stress are 3×10^4 kips/in.^{2}, 0.2836 lb/in.^{3}, and 36.0 kips/in.^{2}, respectively. The constants a_i, c_i, and d_i are 0.58, 0.58, and 0.67, respectively. The lower limit on the moment of inertia of each element is 1.0 in.^{4} and there is no upper limit. The resonant frequency for the frame is 15.0 Hz and the displacement limits are 1.0 in. at nodes 2, 3, 4, 5, 6, and 7 in both x_1- and x_2-directions. There are three loading conditions for the structure, and they are as shown on Fig. 5-28.

With Only Stress Constraints Computation time = 21.47 sec			With All the Constraints Computation time = 48.84 sec		
El. No.	Starting Values, in.^{4}	Final Values, in.^{4}	El. No.	Starting Values, in.^{4}	Final Values, in.^{4}
1	6400.0	3264.8	1	6400.0	3794.0
2	6400.0	901.4	2	6400.0	1436.3
3	6400.0	801.5	3	6400.0	845.7
4	6400.0	801.5	4	6400.0	845.7
5	6400.0	2598.7	5	6400.0	4618.8
6	6400.0	2598.7	6	6400.0	4618.8
7	6400.0	901.4	7	6400.0	1436.3
8	6400.0	3267.4	8	6400.0	3794.0
Wt, lb	15790.8	8292.0	Wt, lb	15790.8	9722.5

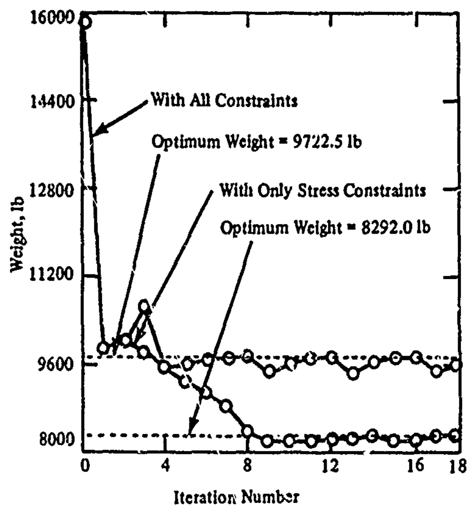


Figure 5-29. Iteration Number vs Weight Curves for Example 5-5; One-bay, Two-story Frame

the first few iterations that reductions in the values of the design variables for elements 23, 24, 25, 28, 29, and 30 were relatively smaller than those of other elements. This is due to nature of the gradient of objective function for this problem (Eq. 5-106). So, the values of these design variables were reduced considerably at the 7th iteration. This is shown by the vertical drop in the graph at the 7th iteration on Fig. 5-31. In the second case, where all the constraints were considered, variation of the cost function with respect to the iteration is shown in Figure 5-32. In this case, the starting point was infeasible and the convergence was obtained in 8 iterations.

tion for the problem is given in Table 5-14. The frame was first optimized by imposing the stress constraint only. The optimum weight in this case was 21706.6 lb with a computation time of 8.32 min for 21 iterations. At the final design point, the maximum constraint violation was 0.025% for stress in element number 25. Next, the frame was designed by imposing all the constraints. The optimum weight in this case was 24290.1 lb with a computation time of 8.7 min for 20 iterations. At the final design point, the maximum constraint violation was 0.0072% for displacement in the x_1-direction at node 1. All other violations were less than that.

Fig. 5-31 shows variation of the cost function with respect to the iteration number. The starting point in this case was quite a distance away from the optimum point. Therefore, a larger step size was used in the first few iterations. Also, it was observed from

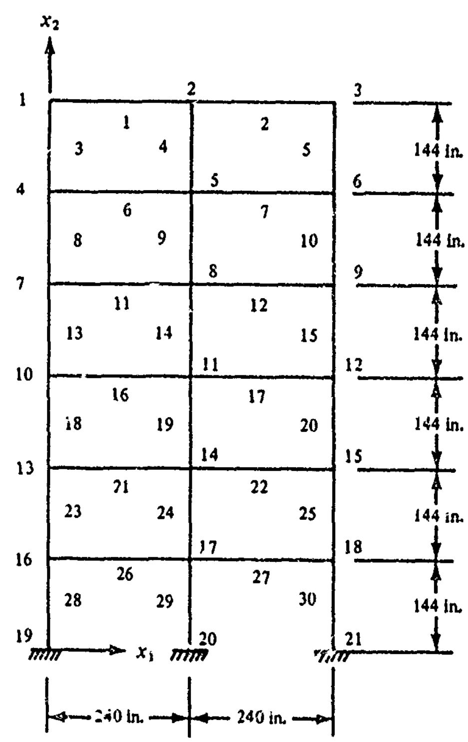


Figure 5-30. Two-bay, Six-story Frame (Example 5-6)

TABLE 5-14

TWO-BAY, SIX-STORY FRAME (EXAMPLE 5-6)

Design Information: For each element of the frame the modulus of elasticity, the specific weight, and the yield stress are 3 x 10^4 kips/in.^{2}, 0.2836 lb/in.^{3}, and 36.0 kips/in.^{2}, respectively. The constants $a_i, c_i,$ and d_i are 0.58, 0.58, and 0.67, respectively. The lower and the upper limits on the moment of inertia of each element are 394.5 in.^{4} and 6699.0 in.^{4}, respectively. The resonant frequency of the structure is taken as 4.0 Hz and the displacement limits are 2.0 in. at all nodes in both x_1- and x_2-directions. There are four loading conditions for the frame: (1) Uniformly distributed load of −4.0 kips/ft on element 1, 7, 11, 17, 21, and 27, and − 1.0 kip/ft on elements 2, 6, 12, 16, 22, and 26; (2) Uniformly distributed load of − 4.0 kips/ft on elements 2, 6, 16, 22, and 26, and − 1.0 kip/ft on elements 1, 7, 11, 17, 21, and 27; (3) Uniformly distributed load of − 1.0 kip/ft on elements 1, 2, 6, 7, 11, 12, 16, 17, 21, 22, 26, and 27, and loads of 9.0 kips each at nodes 1, 4, 7, 10, 13, and 16 in direction of the x-axis; (4) Uniformly distributed load of − 1.0 kip/ft on elements 1, 2, 6, 7, 11, 12, 16, 17, 21, 22, 26, and 27, and loads of − 9.0 kips each at nodes 3,6, 9, 12, 15, and 18 in direction of x_1 - axis.

With Only Stress Constraints Computation time = 8.32 min			With All Constraints Computation time = 8.70 min		
El. No.	Starting Values, in.^{4}	Final Values, in.^{4}	El. No.	Starting Values, in.^{4}	Final Values, in.^{4}
1	2400.0	450.6	1	394.5	473.8
2	2400.0	450.6	2	394.5	473.8
3	2400.0	498.6	3	394.5	467.2
4	2400.0	394.9	4	394.5	437.5
5	2400.0	498.6	5	394.5	467.2
6	2400.0	530.8	6	394.5	568.7
7	2400.0	530.8	7	394.5	569.1
8	2400.0	394.3	8	394.5	394.5
9	2400.0	397.1	9	394.5	608.5
10	2400.0	394.1	10	394.5	394.5
11	3200.0	481.8	11	450.0	787.0
12	3200.0	481.7	12	450.0	786.4
13	3200.0	425.3	13	400.0	412.7
14	3200.0	472.7	14	450.0	794.1
15	3200.0	425.3	15	400.0	412.6
16	4000.0	521.9	16	550.0	930.2
17	4000.0	521.7	17	550.0	930.0
18	4000.0	468.3	18	550.0	561.9
19	4000.0	723.5	19	750.0	920.4
20	4000.0	467.5	20	550.0	561.8
21	4800.0	699.1	21	600.0	1019.1
22	4800.0	699.1	22	600.0	1018.7
23	4800.0	646.3	23	700.0	693.4
24	4800.0	1044.5	24	1100.0	1197.0
25	4800.0	646.5	25	700.0	693.3
26	5600.0	666.4	26	600.0	868.5
27	5600.0	666.4	27	600.0	867.9
28	5600.0	1099.0	28	1200.0	1245.4
29	5600.0	1489.7	29	1600.0	1658.6
30	5600.0	1099.0	30	1200.0	1245.3
Wt, lb	54290.9	21706.6	Wt, lb	21243.6	24290.1

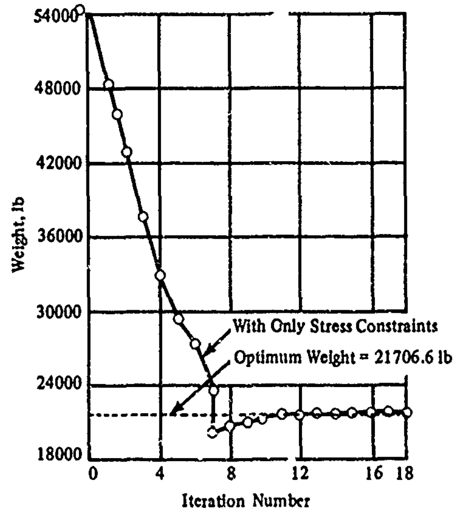


Figure 5-31. Iteration Number vs Weight Curves for Example 5-6; Two-bay, Six-story Frame, With Stress Constraints Only

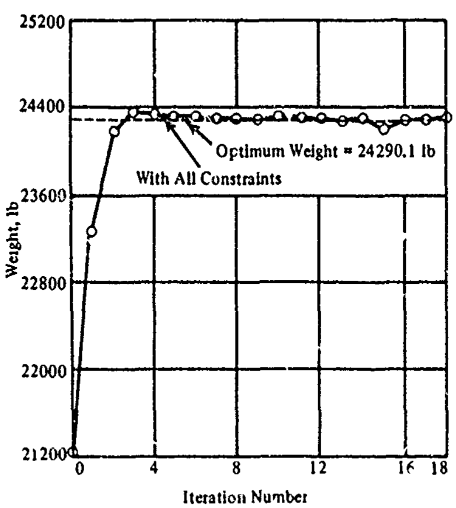


Figure 5-32. Iteration vs Weight Curves for Example 5-6; Two-bay, Six-story Frame, With All Constraints

5-7 INTERACTIVE COMPUTING IN STRUCTURAL OPTIMIZATION

5-7.1 THE INTERACTIVE APPROACH

Structural optimization techniques treated thus far consist of methods which seek to determine an optimum design, within a well-defined mathematical structure, by purely mathematical techniques. A second approach consists of providing the designer with an interactive computing tool with which he can try nominal designs, get rapid analysis feedback, and alter his initial design based on his knowledge of structural behavior. Both methods have been used with varying degrees of success on a variety of design problems. In general, the first approach has been used for problems with well-defined optimality criteria, such as minimum weight or maximum stiffness. The second approach has been used to aid designers in large scale structural design problems, primarily airframe design, such as the Air Force C-5 transport aircraft.

The possibility of utilizing a combination of these two methods for structural design has been the subject of a recent paper (Ref. 36). This paragraph presents the specifics of application of the steepest-descent technique with designer interaction. This hybrid approach is appealing from a number of points of view. First, the problem of topological design, i.e., determination of optimum structural configuration, has been addressed with very limited success from an analytical point of view. Topological design, in practice, is done by experienced structural designers, occasionally with the aid of interactive computation. Combined analytical and interactive computing methods appear to be essential for this important class of problems. A second problem area arises due to the difficulty in formulating a single optimality condition and mathematically precise design constraints. Often, conflicting design constraints and objectives arise during design which require experienced judgment and defy *a priori* mathematical formulation. Such problems appear to require an interactive computing capability but should profit from analytical methods that are used in automated structural optimization.

Due to unavailability of a large scale, interactive system, the computations for this study were simulated. Instructions were prepared and computations were run in the batch mode. Output data were then displayed and analyzed just as they would be in the interactive mode, and instructions for recomputation were given by the designer and the process repeated. The delay in designer interaction is felt to degrade performance somewhat, over true interactive computing, since the designer tends to forget pertinent detailed data during the time delay. For this reason, the results of this study should provide a conservative estimate of the designer's performance in a truly interactive mode.

5-7.2 INTERACTIVE STRUCTURAL DESIGN USING SENSITIVITY DATA

The steepest-descent optimization method developed in this Chapter has been used to solve a number of relatively large scale structural optimization problems with good success. All these problems, however, have been well formulated mathematically and have involved structures with a predetermined form. Difficulties have occurred when certain structural elements tend toward zero cross section. Further, no universal method has been found to determine the best step-size η in the optimization algorithm. These and other inherent difficulties in automated optimization lead one to interject an experienced designer into the computational, optimization algorithm. The result is a hybrid structural optimization technique.

Reconsidering the design improvement step of the optimization algorithm, one might draw a vector picture in design space, as is depicted in Fig. 5-33. Here, $-\eta\delta b^1$ is the direction which will yield the greatest reduction in J subject to the required constraints, and δb^2 is the design change required to give the desired constraint error correction. While useful in this form, there is a better display of these data for use by the experienced structural designer. The scalar components of $-\delta b^1$ and δb^2 tell the designer whether he should increase or decrease his individual design variables to obtain desirable changes in overall structural response. Further, relative importance of design variable changes is given. For this reason, δb^1 may be interpreted as a vector of design sensitivity coefficients that relate individual design parameter changes to overall structural characteristics. It is extremely important to note, at this point, that these sensitivity coefficients account for constraints implicitly; i.e., the direction of change indicated in the

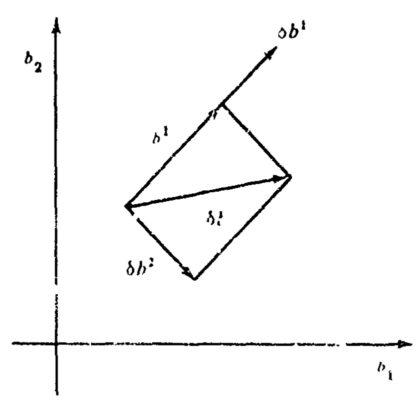


Figure 5-33 Vector Change in Design Space

design parameters will not cause significant violation in specified performance constraints such as stress limits and deflection limits.

To illustrate these ideas, consider the simple structural design problem in Fig. 5-34.

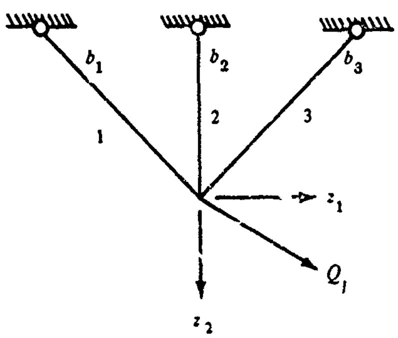


Figure 5-34. Three-bar Truss

The cost function here is structural weight. If, for example, the stress in member 1 is at its allowable limit under one of the loads, then the indicated changes in design $(-\delta b_1^1, -\delta b_2^1, -\delta b_3^1)$ will not increase the stress in member 1. To make the design sensitivity data of maximum use to the designer, consider the graphical display in Fig. 5-35. In this display, σ_i are the stresses in the various members. This display gives the experienced designer a clear picture of the manner in which he should change his design parameters to reduce total weight, subject to stress constraints. He can now choose the desired reduction ΔJ in weight and take the resulting design change δb, or if he wishes, he can input modified design changes through an interactive computer terminal.

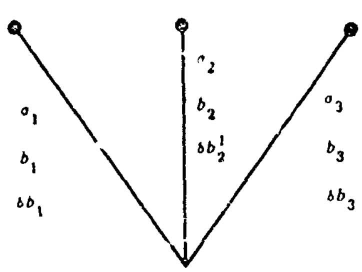


Figure 5-35 Display of Design Sensitivity Data

T a number of other respects in whic. mode of designer interaction with the computer alogrithm is beneficial. First, it ppens in the automated use of the gorithm that oscillation of admissible designs oc urs because too large a design improv t has been requested. Such oscillation can often be identified by the designer after only a few iterations and the step size can be reduced to prevent loss of computer time, which can be significant in large scale problems. Conversely, if an estimate quite far from the optimum is chosen to initiate the algorithm, it often happens that the designer chooses far too small a step size. The result is a very small improvement in the design which can be sensed by the designer and improved before excessive computation time is expended.

A second important benefit from designer interaction with the algorithm arises due to the occurrence of local minima and singularities in the analytical formulation of the design problem The problem of local minima is illustrated by Fig. 5-36. Virtually all optimization methods seek local optim. and do not solve the global optimization problem It is easy for an optimization technique to get hung up at point B and not get to point A, which is the global minima, so the designer must try different starting points to obtain the global solution This is a very time consuming and indefinite technique with very few analytical aids to the designer Part of the

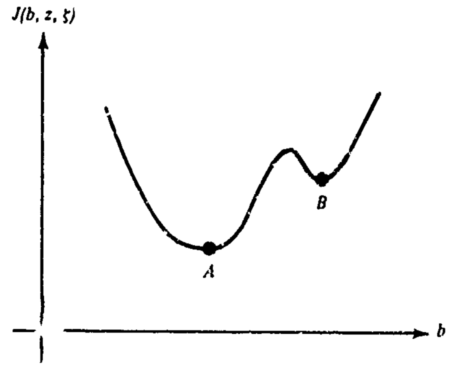


Figure 5-36. Local Optima

difficulty here arises because Figure 5-36 is the wrong display for the designer in that it does not utilize his knowledge and experience with structures.

A much better approach for the designer is to look at a display such as Figure 5 35. He can use his experience to restart the optimization algorithm at a meaningful distribution of design variables which may be quite different from the design which resulted from previous calculations. His experience, thus, aids him in starting with different trial designs.

Perhaps even more important than trying various distributions of design variables, the designer can utilize the display of Fig. 5-35 to change the configuration of the structure based on information he accumulates during iterative design and based on his experience For example he might try taking member 2 out of the structure and optimize based on the modified configuration. Very often, significant gains are made in this manner Precisely this behavior occurs in the three member truss being considered

There are actually compelling mathematical reasons for allowing the designer to make changes in configuration as outlined. There are no general optimization methods, to date, which will remove a member during iterative design. The reason is that as a member cross section goes toward zero, as is required to remove a member, the equations of structural mechanics and stress constraints become singular. This sort of behavior is typical when the configuration of a system is changed and a different set of equations is required to describe the behavior. At the present time, allowing the designer to make changes in configuration appears to be the most feasible approach, which requires that he play an active role in the iterative optimization algorithm.

5-7.3 EXAMPLE PROBLEMS

1. Example 5-7. A Three-member Truss

As an illustrative example of the technique presented in par. 5-7.2, an elementary optimal design problem will be solved under a number of loading conditions and a variety of constraints The effect of designer-computer interaction on rate of convergence is examined as well as the effect of changing structural configuration.

Figure 5-37(A), shows the geometry and dimensions of the structure being considered. This structure has been studied by Schmit (Ref. 37), Sved and Ginos (Ref. 38), and Corcoran (Ref. 35). Three independent loading conditions are applied to the structure These are as follows. 40K at 45 deg, 30K at 90 deg, 20K at 135 deg. The allowable stress level for members 1 and 3 is ± 5 Ksi and for member 2 it is ± 20 Ksi The density of the material is taken as 0 10 lb/in ^{3} and Young's modulus as 10^4 Ksi Starting from the feasible solution, b_1 - 8 0, b_2 = 2 4, b_3 = 3 2, Schmit (Ref 37) arrived at the solution b_1 = 7 099,

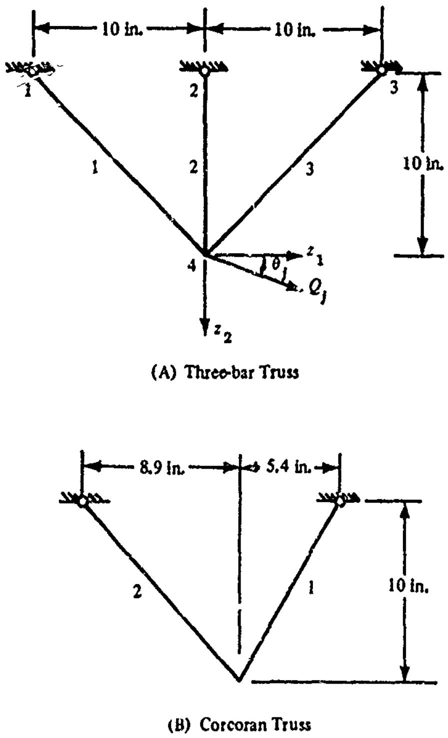


Figure 5-37. Trusses (Example 5-7)

b_2 = 1.849, b_3 = 2.897, for which J = 15.986 lb. Sved and Ginos (Ref. 38) have shown that this is only a local minima and by omitting member 3, they obtained the solution as b_1 = 8.5, b_2 = 1.5 with J = 12.812 lb. They have also shown that it is impossible to reach this minimum by an iterative optimization method unless member 3 is omitted from the calculations by the designer. Corcoran (Ref. 35) has considered configurational optimization of this three-bar truss. By using horizontal coordinates of nodes 1, 2, and 3 also as design variables, he arrived at an optimum structure shown in Fig. 5-37(B). As a result of this configurational optimization procedure, members 1 and 3 were combined and their orientation is shown by member 1 of Fig 5-37(B). Member 2 attained an orientation as shown in this figure. The final solution obtained by Corcoran was b_1 = 4.241, b_2 = 2.038 with J = 7.55 lb.

Considerable experimentation was done with this problem. Starting from a feasible point b_1 = 10, b_2 = 5, b_3 = 5, the solution obtained without interaction was b_1 = 7.064, b_2 = 1.971, b_3 = 2.635 and the minimum was J = 15.97 lb. The variation of weight with respect to iteration number is shown by Curve 1, Fig. 5-38. Next, by adjusting the step size in interactive computing, the solution was obtained in only five iterations. This is shown by Curve 2, Fig. 5-38. It was observed that member 2 never reached its allowable stress level. As a second starting point, the area of member 2 was initially chosen to bring its stress to the allowable limit. The minimum reached in this case was the same as before, Curve 3, Fig. 5-38. Another solution was obtained by starting from an infeasible point b_1 = 5.0, b_2 = 1.5, b_3 = 0.10. The solution in this case was b_1 = 6.98, b_2 = 2.30, b_3 = 2.68 with J = 15.97 lb, Curve 4, Fig 5-38.

Next, member 3 was omitted from the structure. Starting from a point b_1 = 10, b_2 = 5, the solution obtained was b_1 = 8.0, b_2 = 1.5 with J = 12.812 lb, Curve 1, Fig. 5-39, which is same as reported in Ref 38. At

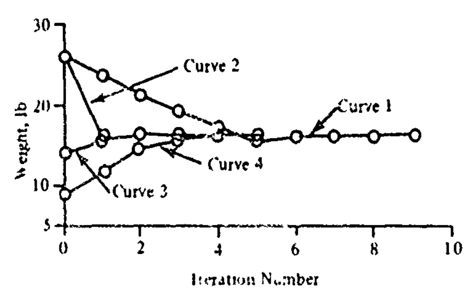


Figure 5-38. Iteration vs Weight Curves for Example 5-7, Three-bar Truss With Stress Constraints Only

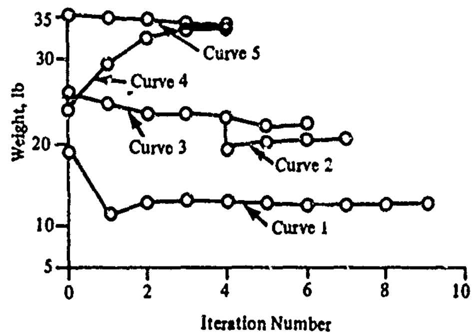


Figure 5-39. Iteration vs Weight Curves for Example 5-7, Three-bar Truss With All Constraints

this point an interesting observation was made. The maximum horizontal and the vertical deflections of node 4 were as follows: with three bars, $z_1 = 0.689 \times 10^{-2}$ in., $z_2 = 0.595 \times 10^{-2}$ in.; with two bars, $z_1 = 0.239 \times 10^{-1}$ in., and $z_2 = 0.20 \times 10^{-1}$ in. Thus, although the optimum weight obtained by omitting member 3 is approximately 24% lower than the weight obtained by including member 3, the deflections of node 4 in the former case were approximately four times greater than in the latter case.

One might be led to believe that if deflection or frequency constraints were enforced, then the optimum structure might not be statically determinate. To investigate this possibility, displacement as well as buckling and natural frequency constraints were imposed. The deflection limits were taken as $z_1 = \pm 0.005$ in. and $z_2 = \pm 0.005$, and the lower limit on natural frequency was taken as 3830 Hz. With the starting point $b_1 = 10$, $b_2 = 5$, $b_3 = 5$, the solution obtained was $b_1 = 9.18$, $b_2 = 2.16$, $b_3 = 3.85$, and $J = 20.59$ lb, Curves 2 and 3, Fig 5-39. When member 3 was omitted, the starting point was taken as $b_1 = 10$, $b_2 = 10$ Curve 4, Fig 5-39, and as $b_1 = 18$, $b_2 = 10$, Curve 5, Fig 5-39. The solution obtained in this case was $b_1 = 16.0$, $b_2 = 11.31$, and $J = 33.94$ lb. Thus, the optimum weight obtained for this statically determinate case is approximately 70% higher than the optimum weight obtained for the statically indeterminate case.

It was found that interactive computing yielded convergence more rapidly than was the case in the batch mode. It is expected that even more significant reduction in computing time will occur in large scale problems.

This problem was also solved by omitting member 2 from computation. The results obtained in this case are given in Columns 3 and 7 of Table 5-15. The truss optimized by Corcoran (Ref. 35) was also solved here by first imposing the stress constraints only and then by considering all the constraints. The results of these cases are given in Columns 4 and 8 of Table 5-15.

The key point in the solution is that the configuration of the optimum design is not obvious from analytical considerations. A designer's experience and insight are required to select candidate configurations and then obtain the optimum design analytically. The global solution in this case must be chosen by comparing relative minima. It may be expected, in structures with greater redundancy, that certain members may be removed during interactive computation when they are observed to approach their allowable lower limits.

An interesting point, illustrated by Table 5-15, is that a statically determinate truss is optimum when only stress constraints are imposed. Quite the contrary, when the full range of constraints are imposed, a statically indeterminate truss is optimum (not considering the configurational optimization).

TABLE 5-15

OPTIMUM THREE-MEMBER TRUSSES (EXAMPLE 5-7)

El No.	With Only Stress Constraints Final Area, in.^{2}				With All Constraints Final Area, in.^{2}			
	1	2	3	4	5	6	7	8
1	7.064	8.500	7.991	Corcoran Truss	9.180	16.00	8.485	Corcoran Truss
2	1.971	1.500	–	4.246	2.160	11.310	–	4.247
3	2.835	–	4.243	2.039	3.850	–	8.485	11.410
Wt, lb	15.970	12.812	17.300	7.555	20.59	33.94	24.000	20.115
Max. Defl, in.	0.00689	0.02390	0.00766	0.02559	0.005	0.005	0.005	0.005

2. *Example 5-8. Transmission Tower*

Fig. 5-40 shows the geometry and dimensions of the transmission tower to be studied. This problem has been considered by Venkayya and others (Ref. 39). The tower has 25 members, 10 joints, 18 degrees of freedom, and is designed for 6 loading conditions. The structure is indeterminate, with a degree of indeterminacy of seven.

The tower was designed by first imposing only stress constraints, and then by imposing stress, displacement, buckling, and natural frequency constraints. Design information is given in Table 5-16, and the final results obtained are shown in Tables 5-17 and 5-18. Table 5-17 shows the results when only stress constraints are considered, and Table 5-18 gives the results for the corresponding cases when all the constraints are considered. For results given in Column 1 of Table 5-17, all the members of tower were included in the computation and the Curve 1 of Fig. 5-41 shows the variation of cost function with the number of iterations. The computations of this case were monitored to determine which cross sections went to their lower bounds.

One set of members which attained their lower limits of cross-sectional area were numbers 10, 11, 12, and 13. It was observed that these members carried small forces and could be removed without causing collapse of the tower, so they were removed from the tower. The final values of areas of cross section of the resulting structure are given in Column 2 of Table 5-17. Curve 2 of Fig. 5-41 shows the variation of cost function with respect to the design cycle. The final weight in this case was slightly less than the previous case.

The next member that reached its lower limit was number 1, so it was also removed from the structure. The results of this case are given in Column 3 of Table 5-17 and Curve 3

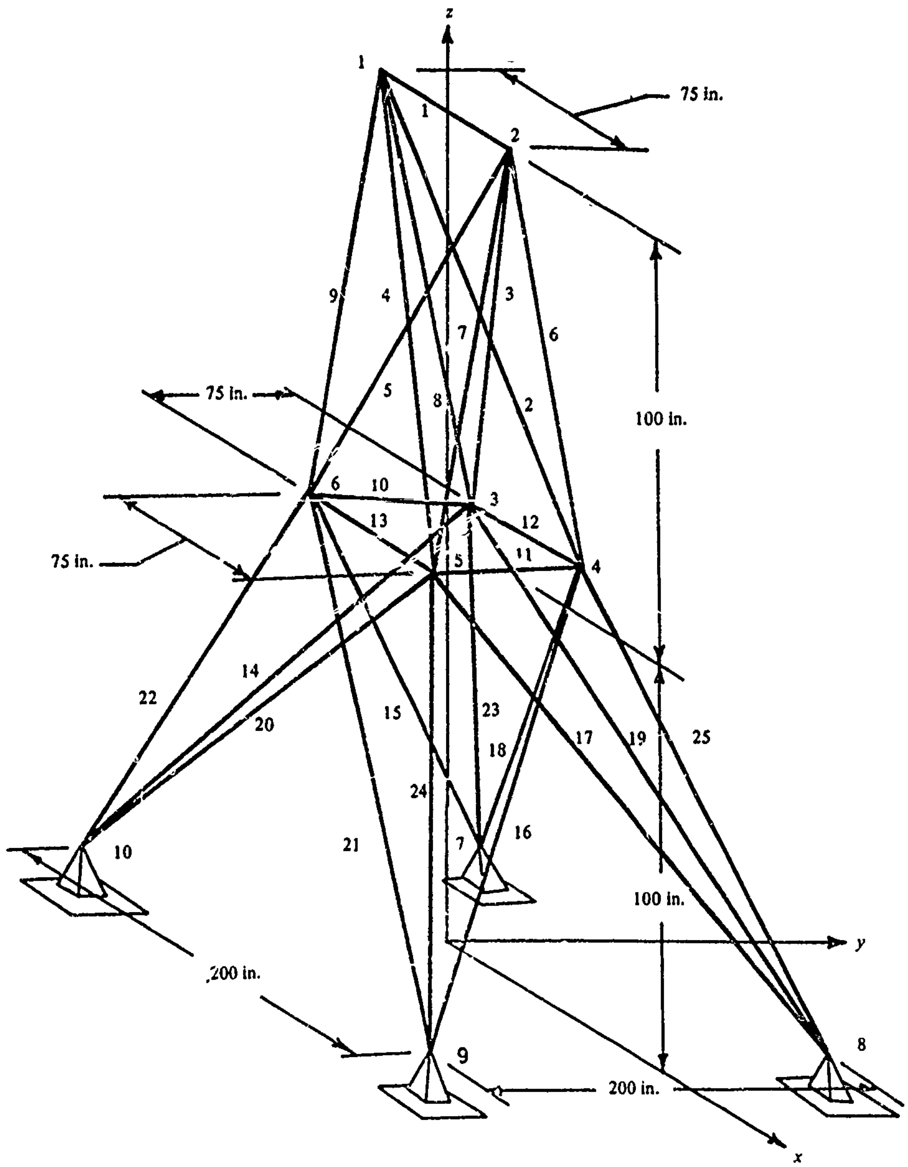


Figure 5-40. Transmission Tower (Example 5-8)

TABLE 5-16

DESIGN INFORMATION FOR TRANSMISSION TOWER (EXAMPLE 5-8)

For each member of the structure, the modulus of elasticity E_i, the specific weight ρ_i, the constant α_i (moment of inertia of ith member, $I_i = \alpha_i b_i^2$), and the stress limits are 10^4 kips/in.^{2}, 0.10 lb/in.^{3}, 1.0, and ± 40.0 kips/in.^{2}, respectively. The lower limit on the area of cross section of each member is 0.10 in.^{2} for the case with stress constraints only and 0.01 in.^{2} for other cases. There is no upper limit on the member sizes. The resonant frequency for the structure is 173.92 Hz and the displacement limits are 0.35 in. on all nodes and in all directions. There are six loading conditions and they are as follows (all loads are in kips):

Load Cond.	Node	Direction of Load			Load Cond.	Node	Direction of Load		
		x	y	z			x	y	z
1	1	1.0	10.0	−5.0	2	1	0	10.0	−5.0
	2	0	10.0	−5.0		2	−1.0	10.0	−5.0
	3	0.5	0	0		4	−0.5	0	0
	6	0.5	0	0		5	−0.5	0	0
3	1	1.0	−10.0	−5.0	4	1	0	−10.0	−5.0
	2	0	−10.0	−5.0		2	−1.0	−10.0	−5.0
	3	0.5	0	0		4	−0.5	0	0
	6	0.5	0	0		5	−0.5	0	0
5	1	0	20.0	−5.0	6	1	0	−20.0	−5.0
	2	0	−20.0	−5.0		2	0	20.0	−5.0

of Fig. 5-41. The final weight in this case was 86.94 lb, which is given slightly less than the previous case. Finally, members 14, 15, 16, and 17 were at their lower limits of cross-sectional area. Removal of any of these members, however, would cause collapse of the structure. Members 2 and 5 or 3 and 4 could be removed to make the structure determinate. The results for a statically determinate structure, obtained by removing members 2 and 5, are shown in Column 4 of Table 5-17. The final weight in this case was 106.97 lb. It may be noted that this statically determinate structure yielded only a local optimum, Curve [illegible], Fig. 5-41.

Another sequence of removing the members that reached their lower limits of area of cross section was also tried. Members 14, 15, 16, and 17 reached their lower bounds but removal of all of these members rendered a structure that was geometrically unstable. However, members 14 and 16 or 15 and 17 could be removed without causing the collapse of the structure. Results with members 15 and 17 removed are given in Column 5 of Table 5-17 and similar results are obtained by omitting members 14 and 16 from the computation. The next set of members that were at their lower bounds and could be removed without making the structure unstable were numbers 1, 12, and 13. These were also removed from the structure and the results obtained in this case are given in Column 6 of Table 5-17. Two other members could be removed from the structure to make it statically determinate. Results obtained by removing members 4 and 5, and then numbers

TABLE 5-17

OPTIMUM TRANSMISSION TOWERS WITH STRESS CONSTRAINTS ONLY (EXAMPLE 5-8)

El. No.	Final Area, in.^{2}							
	1	2	3	4	5	6	7	8
1	0.100	0.100	–	–	0.100	–	–	–
2	0.376	0.377	0.346	–	0.384	0.364	0.272	–
3	0.376	0.377	0.346	0.100	0.384	0.366	0.272	0.272
4	0.376	0.377	0.346	0.100	0.387	0.363	–	0.272
5	0.376	0.377	0.346	–	0.385	0.365	–	–
6	0.471	0.470	0.494	0.779	0.465	0.484	0.775	0.779
7	0.471	0.470	0.494	0.779	0.463	0.482	0.779	0.775
8	0.471	0.470	0.494	0.779	0.464	0.481	0.779	0.779
9	0.471	0.470	0.494	0.779	0.463	0.479	0.779	0.779
10	0.100	–	–	–	0.103	0.103	0.182	0.182
11	0.100	–	–	–	0.103	0.103	0.182	0.182
12	0.100	–	–	–	0.100	–	–	–
13	0.100	–	–	–	0.100	–	–	–
14	0.100	0.100	0.100	0.165	0.151	0.152	0.302	0.302
15	0.100	0.100	0.100	0.165	–	–	–	–
16	0.100	0.100	0.100	0.165	0.151	0.152	0.302	0.302
17	0.100	0.100	0.100	0.165	–	–	–	–
18	0.277	0.279	0.292	0.413	0.278	0.288	0.413	0.413
19	0.277	0.279	0.292	0.413	0.277	0.288	0.413	0.413
20	0.277	0.279	0.292	0.413	0.274	0.287	0.413	0.413
21	0.277	0.279	0.292	0.413	0.273	0.287	0.413	0.412
22	0.380	0.374	0.363	0.547	0.445	0.436	0.669	0.669
23	0.380	0.374	0.363	0.547	0.334	0.370	0.447	0.447
24	0.380	0.374	0.363	0.547	0.442	0.436	0.669	0.669
25	0.380	0.374	0.363	0.547	0.336	0.370	0.447	0.447
Wt, lb	91.13	87.90	86.94	106.97	89.94	88.95	113.69	113.68
Max. Defl. in.	2.288	2.305	2.311	3.489	2.486	2.453	3.614	3.615

2 and 5 are given, respectively, in Columns 7 and 8 of Table 5-17. Computations were also carried out by removing members 2 and 3, and members 3 and 4 along with members 1, 12, 13, 15, and 17. Results obtained in these cases were the same as those shown in Columns 7 and 8 of Table 5-17. For this reason, these results are not reproduced here. Finally, another determinate structure, obtained by removing members 1, 2, 5, 15, 16, 19, and 20 was optimized. The cross-sectional areas of various members at the optimum point were as follows. 3,4(0.100), 6 to 9(0.779); 10,11(0.182), 12 13(0.446), 14,17(0 302), 18,21(0 775), 22,25(0 537), and 23,24(0.751). The optimum weight in this case was 118 1 lb and the maximum deflection at this point was 3.861 in

TABLE 5-18

OPTIMUM TRANSMISSION TOWERS WITH ALL CONSTRAINTS
(EXAMPLE 5-8)

El. No.	Final Area, in.^{2}							
	1	2	3	4	5	6	7	8
1	0.010	0.010	–	–	0.010	–	–	–
2	2.092	2.339	2.393	–	2.263	2.389	0.548	–
3	2.075	2.386	2.404	0.548	2.264	2.384	0.548	0.548
4	2.095	2.339	2.393	0.548	2.021	1.826	–	0.548
5	2.083	2.385	2.404	–	1.920	1.915	–	–
6	2.357	2.085	2.076	7.132	2.389	2.452	6.596	6.669
7	2.354	2.084	2.076	6.857	2.186	2.042	6.483	6.296
8	2.350	2.113	2.083	6.895	2.411	2.430	6.596	6.686
9	2.335	2.112	2.082	7.101	2.095	2.123	6.476	6.471
10	0.035	–	–	–	0.668	0.621	2.102	2.054
11	0.035	–	–	–	0.658	0.630	2.102	2.047
12	0.087	–	–	–	0.090	–	–	–
13	0.084	–	–	–	0.071	–	–	–
14	1.113	1.114	1.139	1.785	1.461	1.485	4.172	4.101
15	1.113	1.114	1.139	1.735	–	–	–	–
16	1.112	1.117	1.146	1.727	1.438	1.498	4.170	4.167
17	1.112	1.117	1.146	1.798	–	–	–	–
18	2.056	2.047	2.027	4.317	2.161	2.171	4.692	4.645
19	2.058	2.034	2.022	4.390	2.156	2.173	4.692	4.664
20	2.046	2.047	2.027	4.400	2.403	2.538	4.985	5.108
21	2.058	2.034	2.022	4.328	2.415	2.524	4.989	5.038
22	2.822	2.878	2.886	5.655	4.187	4.035	6.746	6.909
23	2.808	2.878	2.886	5.730	2.915	2.873	5.086	4.781
24	2.803	2.926	2.895	5.743	4.124	4.086	6.743	7.039
25	2.785	2.926	2.895	5.648	2.908	2.881	5.086	4.749
Wt, lb	590.32	596.64	597.82	1060.6	625.37	626.70	1142.7	1139.9
Max. Defl. in.	0.350	0.350	0.350	0.350	0.350	0.350	0.350	0.350

All these tower configurations were also optimized by imposing all constraints, i.e., stress, displacement, buckling, and natural frequency The results of these cases are given in Table 5-18. Curves 1 to 4 of Fig. 5-42 show the variation of cost function with respect to the iteration number for results of Columns 1 to 4 of Table 5-18. It can be observed from the results of Table 5-18 that, for the case in which all constraints were imposed, the optimum weight of the tower increased as more redundant members were removed from the structure.

5-7.4 INTERACTIVE COMPUTING CONCLUSIONS

Computing times for this interactive computing approach are considerably shorter than

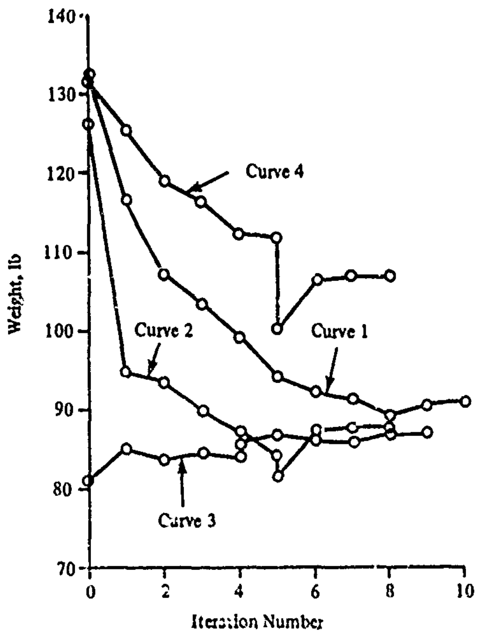


Figure 5-41. Iteration vs Weight Curves for Example 5-8, Transmission Tower With Stress Constraints Only

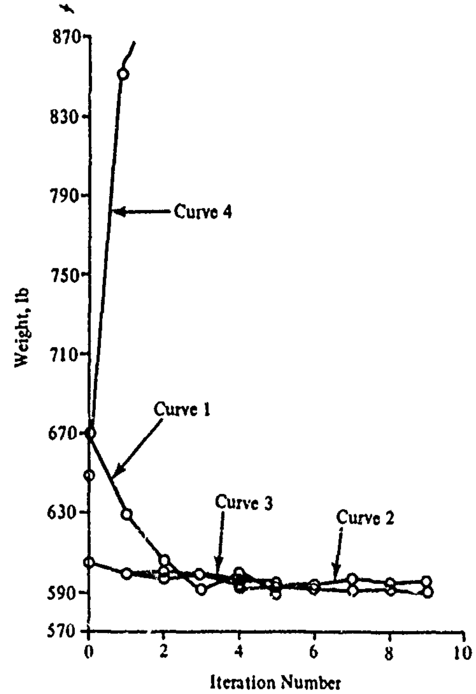


Figure 5-42. Iteration vs Weight Curves for Example 5-8, Transmission Tower With All Constraints

had been experienced when the same problems were solved in the batch mode. Second, and probably more significant, interactive computing allows the designer to alter the structural configuration in a systematic way to seek the global optimum design. This is not to say that a mathematically precise method of obtaining a global optimum has been found, for no such method is known. It appears, however, that the technique presented here makes strong use of the designer's knowledge and intuition, and gives him a tool with which to seek a global optimum in an organized way.

The results presented for the two examples solved in par. 5-7 are of interest in their own right For the case when only stress constraints are imposed, results of Table 5-15 indicate that minimum weight designs for trusses with multiple loading may be statically determinate. However, the results of the second example given in Table 5-17 indicate that all statically determinate trusses may not be lighter than the indeterminate trusses.

For the case when all constraints are imposed, results of Tables 5-15 and 5-18 show that statically indeterminate trusses are lighter than the determinate trusses.

REFERENCES

1. Z. Wasiutynski and A Brandt, "The Present State of Knowledge in the Field of Optimum Design of Structures", *Appl Mech Rev*, Vol 16, No 5 May 1963,

pp. 341-350.

2. C. Y. Sheu and W. Prager, "Recent Developments in Optimal Structural Design", *Appl. Mech. Rev.*, Vol. 21, No. 10, October 1968, pp. 985-992.

3. R. A. Ridha, "Minimum Weight Design of Aircraft Landing Gear Reinforcement Rings", *Proceedings AIAA/ASME 9th Structures, Structural Dynamics and Materials Conference*, April 1968.

4. W. A. Thornton and L. A. Schmit, Jr., "Structural Synthesis of an Ablating Thermostructural Panel", *Proceedings AIAA/ASME 9th Structures, Structural Dynamics and Materials Conference*, April 1968.

5. L. A. Schmit, Jr. and T. P. Kicher, "A Structural Synthesis Capability for Integrally Stiffened Cylindrical Shells", *Proceedings AIAA/ASME 9th Structures, Structural Dynamics and Materials Conference*, April 1968.

6. R. Luik and R. J. Melosh, "An Allocation Procedure for Structural Design", *Proceedings AIAA/ASME 9th Structures, Structural Dynamics and Materials Conference*, April 1968.

7. W. J. Woods and J. H. Sams, III, "Geometric Optimization in the Theory of Structural Synthesis", *Proceedings AIAA/ASME 9th Structure, Structural Dynamics and Materials Conference*, April 1968.

8. R. L. Barnett and P. C. Hermann, *High Performance Structure*, NASA Report, NASA CR-1038, May 1968.

9. M. J. Schrader, *An Algorithm for the Minimum Weight Design of the General Truss*, Division of Solid Mechanics, Structures, and Mechanical Design, Case Western Reserve University, Report No. 24, June 1968.

10. L. P. Felton and M. F. Rubinstein, *Optimal Structural Design*, preprint of paper presented at SAE, Aeronautic and Space Engineering and Manufacturing Meeting, October 1968.

11. G. Sved and Z. Ginos, "Structural Optimization Under Multiple Loading", *Int. J. Mech., Sci.*, Vol. 10, 1968, pp. 803-805.

12. W. A. Thornton and L. A. Schmit, Jr., *The Structural Synthesis of an Ablating Thermostructural Panel*, NASA Report, NASA CR-1215, December 1968.

13. W. M. Morrow, II and L. A. Schmit, Jr., *Structural Synthesis of a Stiffened Cylinder*, NASA Report, NASA CR-1217, December 1968.

14. W. R. Spillers and J. Farrell, "On the Analysis of Structural Design", *J. Math. Anal. and Appl.*, Vol. 25, 1969, pp. 285-295.

15. W. J. Woods, "Substructure Optimization in the Theory of Structural Synthesis", *Proceedings AIAA 7th Aerospace Sciences Meeting*, January 1969.

16. M. J. Turner "Optimization of Structures to Satisfy Flutter Requirements", *Proceedings AIAA Structural Dynamics and Aeroelasticity Specialist Conference*, April 1969.

17. C. P. Rubin, "Dynamic Optimization of Complex Structures" *Proceedings AIAA*

Structural Dynamics and Aeroelasticity Specialist Conference, April 1969.

18. R. L. Fox and M. F. Kapoor, "Structural Optimization in the Dynamics Response Regime: A Computational Approach", *Proceedings AIAA Structural Dynamics and Aeroelasticity Specialist Conference,* April 1969.

19. W. C. Hurty and M. F. Rubenstein, *Dynamics of Structures,* Prentice-Hall, Englewood Cliffs, N. J., 1966.

20. O. C. Zienkiewicz, *The Finite Element Method in Structural and Continuum Mechanics,* McGraw-Hill, New York, 1967.

21. R. E. Beckett and J. Hurt, *Numerical Calculations and Algorithms,* McGraw-Hill, New York, 1967.

22. S. H. Crandall, *Engineering Analysis,* McGraw-Hill, New York, 1956.

23. T. Kato, *Perturbation Theory for Linear Operations,* Springer-Verlag, New York, 1966.

24. J. S. Przemieniecki, *Theory of Matrix Structural Analysis,* McGraw-Hill Book Co. New York, 1968.

25. V. B. Venkayya, N. S. Khot, and V. S. Reddy, *Energy Distribution in an Optimum Structural Design,* Technical Report AFFDL-TR-68-158, Wright-Patterson Air Force Base, Ohio 45433. March 1969.

26. D. Johnson and D. M. Brotton, "Optimum Elastic Design of Redundant Trusses", *Journal of the Structural Division, Proc. ASCE* Vol 95, No ST12, December, 1969, pp 2589-2610

27. *Manual of Steel Construction,* Sixth Edition, American Institute of Steel Construction, 1967.

28. Y. Nakamura, *Optimum Design of Framed Structures Using Linear Programming,* Master's thesis, Department of Civil Engineering, M.I.T., Cambridge, Massachusetts, 1966.

29. D. Kavlie and J. Moe, "Application of Nonlinear Programming to Optimum Grillage Design with Nonconvex Sets of Variables", *International Journal for Numerical Methods in Engineering,* Vol. 1, No. 4, 1969, pp. 351-378.

30. D. M. Brown, and A. H. S. Ang, "Structural Optimization by Nonlinear Programming", *Journal of the Structural Division, Proc. ASCE,* Vol. 92, No. ST6, December, 1966, pp. 319-340 and Vol. 93, No. ST5, October, 1967, pp. 618-619.

31. F. Moses, and S. Onoda, "Minimum Weight Design of Structures With Application to Elastic Grillages", *International Journal for Numerical Methods in Engineering,* Vol. 1, No. 4, 1969, pp. 311-331.

32. W. McGuire, *Steel Structures,* Prentice-Hall Inc. Englewood Cliffs, N. J., 1968.

33. *Proceedings, Army Symposium on Solid Mechanics, 1970 – Lightweight Structures,* Army Materials and Mechanics Research Center, Watertown, Mass., 1970.

34. J. Arora, *Optimal Design of Elastic Structures Under Multiple Constraint Conditions,* Dissertation, University of Iowa, 1971

35. P. Corcoran, "Configurational Optimization of Structures", *International Journal of Mechanical Sciences*, Vol. 12, 1970, pp. 459-462.

36. R. Douty and S. Shore, "Technique for Interactive Computer Graphics in Design", *Journal of the Structural Division, ASCE*, Vol. 97, No. ST1, January 1971, pp. 273-288.

37. L. A. Schmit, "Structural Design by Systematic Synthesis", *Second Conference on Electronic Computation*, Structural Division of ASCE, September 1960, pp. 105-132.

38. G. Sved and Z. Ginos, "Structural Optimization Under Multiple Loading", *International Journal of Mechanical Sciences*, Vol. 10, 1968, pp. 803-805.

39. V. B. Venkayya, N. S. Khot, and V. S. Reddy, *Energy Distribution in an Optimum Structural Design*, Technical Report AFFDL-TR-68-158, Wright-Patterson Air Force Base, Ohio 45433, March 1969.

CHAPTER 6

THE CALCULUS OF VARIATIONS AND OPTIMAL PROCESS THEORY

6-1 INTRODUCTION

The problems of Chapters 2 through 5 are all optimal design problems in which the design variables were elements of R^n, i.e., a vector of n real numbers uniquely specified the design of the system being investigated. In many important, real-world, optimal design problems the design of a system cannot be specified so easily. For example, the thrust vector acting on a rocket during takeoff must be continuously oriented in time so that the rocket remains stable and follows a certain path. In this example, the angles the thrust vector makes with the rocket must be specified at each instant of time during takeoff. It is clear that a function specifies the thrust direction rather than a finite number of parameters.

Examples of this kind of problem abound in the aircraft guidance literature and in the optimal control literature. Typical design or control variables in these problems are thrust, motor torque, control surface settings, etc. All these variables must be specified throughout the entire interval of time an aircraft is in the air. Similar problems arise in the presently developing field of optimal structural design. In this field the design variables are generally variables that describe the distribution of material in structural elements.

In order to illustrate the kind of problem to be treated in this chapter, two classic examples will be given.

Example 6-1: The shortest path between two points, (t^0,x^0) and (t^1,x^1), in the t-x plane is to be found. As shown in Fig. 6-1, the particular path chosen between the two points has a length associated with it. The problem is to choose the curve $\hat{x}(t)$, $t^0 < t < t^1$ which has the shortest length. For a smooth curve $x(t)$ the length is given by

$$J(x) = \int_{t^0}^{t^1} \left[1 + \left(\frac{dx}{dt}\right)^2\right]^{1/2} dt. \qquad (6\text{-}1)$$

Note that in this example the quantity $J(x)$ to be minimized is a real number once the function $x(t)$, $t^0 < t < t^1$ is chosen. In this sense $J(x)$ is a real valued function of a function or curve.

Example 6-2: (The Brachistochrone): Given two points (t^0,x^0) and (t^1,x^1) in a vertical plane that do not lie on the same vertical line, find a curve $x(t)$, $t^0 < t < t^1$, joining them so that a particle starting at rest will traverse the curve without friction from one point to the

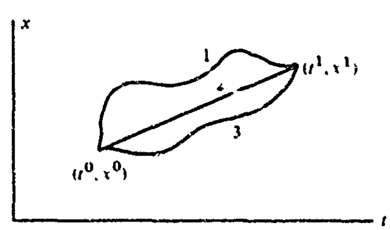


Figure 6-1. Shortest Path

other in the shortest possible time. Candidate curves are shown in Fig. 6-2.

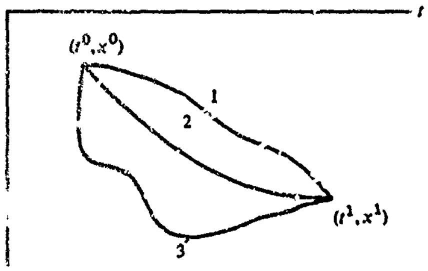


Figure 6-2. Curve for Minimum Time

Let m be the mass of the particle and g be the acceleration due to gravity. Since the particle starts at rest at (t^0,x^0) and there is no friction,

$$\frac{1}{2}mv^2 = mg(x - x^0) \qquad (6\text{-}2)$$

where v is velocity,

$$v = \left[\left(\frac{dt}{d\tau}\right)^2 + \left(\frac{dx}{d\tau}\right)^2\right]^{1/2}$$

$$= \left[1 + \left(\frac{dx}{dt}\right)^2\right]^{1/2} \frac{dt}{d\tau}, \qquad (6\text{-}3)$$

where τ is time. Solving Eq. 6-2 for v substituting this into Eq. 6-3 and solving for $d\tau$ yields

$$d\tau = \frac{\left[1 + \left(\frac{dx}{dt}\right)^2\right]^{1/2} dt}{[2g(x - x^0)]^{1/2}}$$

The total time T required for the particle to move from (t^0,x^0) to (t^1,x^1) is then

$$T = J(x) = \int_{t^0}^{t^1} \frac{\left[1 + \left(\frac{dx}{dt}\right)^2\right]^{1/2}}{[2g(x - x^0)]^{1/2}} dt$$

(6-4)

This notation makes it clear that T depends on the entire curve transversed by the particle. The Brachistochrone problem, therefore, is reduced to finding a curve $\hat{x}(t)$, $t^0 < t < t^1$, that passes through the two given points and makes T as small as possible.

In Examples 6-1 and 6-2 it is clear that a curve, or equivalently a function characterizing the curve, is to be found as the solution of the optimization problem. Further, the real valued quantities to be minimized are determined by curves or the functions characterizing those curves. These real valued quantities, therefore, are functions of functions. Such a real valued function is called a functional. The functional notation $J(x)$ in Eqs. 6-1 and 6-4 is then interpreted as a real valued function of the function $x(t), t^0 < t < t^1$. The most common kind of functional encountered in calculus of variations is the integral.

The optimization problem considered here might be stated as: find the function $x(t)$, $t^0 < t < t^1$, that minimizes the functional $J(x)$. A glance at the functionals defined in Eqs. 6-1 and 6-4 reveals a basic flaw in this statement of the optimization problem. In both cases, the functionals are defined only if the function $x(t)$ has an integrable derivative on $t^0 < t < t^1$, i.e., it doesn't make sense to admit all functions as candidates for an extremum. The problem is more reasonably stated: find the function $x(t)$, $t^0 < t < t^1$, in a class of functions D, that minimizes the functional

$J(x)$. The admissible class of functions here plays a role similar to the constraint sets of Chapters 3, 4, and 5.

The idea of classes of functions required here is basic to the mathematical field called Functional Analysis. Classes of functions in this field are called function spaces. Consider, for example, the collection of all continuous functions $x(t)$ on $0 < t < 1$. The graphs of several such functions are shown in Fig. 6-3.

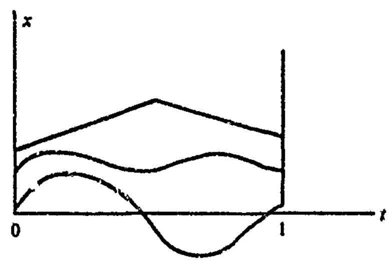


Figure 6-3. Examples of Continuous Functions

It is clear that there are infinitely many continuous functions but that not all functions are contained in this class. For example

$$x(t) = \begin{cases} 0, & 0 < t < 1/2 \\ 1, & 1/2 < t < 1 \end{cases}$$

is not continuous so it is not in the class.

To expedite the development that follows, some notation will be introduced. The collection of continuous functions on $0 < t < 1$ described previously is called a function space and is denoted

$$C^0(0,1) = \{x(t), 0 < t < 1 \mid x(t) \text{ is continuous}\}. \qquad (6\text{-}5)$$

A large number of important function spaces may be described in a similar manner as

$$C^l(a,b) = \{x(t), a < t < b \mid x(t) \text{ has } l \text{ continuous derivatives}\}. \qquad (6\text{-}6)$$

It should be understood here that $x(t)$ may be a vector valued function and the differentiability requirement in Eq. 6-6 refers to each component.

Function spaces may be thought of as sets of elements, where elements in the function space are really curves or functions. In this way the problem of minimizing $J(x)$ may be viewed as picking the element (curve) in the appropriate function space that makes $J(x)$ as small as possible. This approach makes minimization of a functional sound very similar to the programming problems of Chapter 2. With this mental analogy one may begin his study of the calculus of variations armed with a powerful intuitive tool.

The basic ideas of function space theory are presented very clearly in Ref. 1, Chapter 2.

In connection with vector spaces, it is often necessary to require that a function is small, or near the zero function. For this purpose it is required that size of a function be defined. This is done by defining a norm as a functional $\|x\|$ on the function space of interest with the following properties:

$$\|x\| > 0, \|x\| = 0 \text{ implies } x \text{ is the zero function} \qquad (6\text{-}7)$$

$$\|\alpha x\| = |\alpha| \ \|x\| \text{ for real } \alpha \qquad (6\text{-}8)$$

$$\|x + y\| \leq \|x\| + \|y\| . \qquad (6\text{-}9)$$

Examples of norms include

$$\| x \| = \max_{t^0 \le t \le t^1} | x(t) | \qquad (6\text{-}10)$$

for $x \in C^0(t^0, t^1)$ and

$$\| x \| = \left[\int_{t^0}^{t^1} x^2(t)\, dt \right]^{1/2} \qquad (6\text{-}11)$$

for square integrable functions $x(t)$. For a discussion of the basic ideas of functional analysis as they apply to optimization theory the reader is referred to Ref. 1.

With the idea of norm defined, one can speak of relative minima of functionals. The functional $J(x)$ has a relative minimum at $\hat{x} \in D$ if there is a $\delta > 0$ such that

$$J(\hat{x}) \le J(x)$$

for all $x \in D$ with

$$\| \hat{x} - x \| < \delta . \qquad (6\text{-}12)$$

This simply says that $J(x)$ has a minimum in a sufficiently small neighborhood of $\hat{x}$. It is interesting to look at a neighborhood of a curve in $C^0(t^0, t^1)$ where norm is defined by Eq. 6-10. In this case Eq. 6-12 simply demands that $x(t)$ be within δ of $\hat{x}(t)$ for all t in $t^0 \le t \le t^1$. The neighborhood of $\hat{x}$ in this case is simply the collection of all continuous curves which can be drawn between $\hat{x}(t) + \delta$ and $\hat{x}(t) - \delta$, as shown in Fig. 6-4.

The present chapter will be devoted almost exclusively to the theory of the calculus of variations and optimal process theory. Constructive methods for these problems will be treated in the chapters to follow. A knowledge of this basic theory is essential for successful application of the theory of optimal design. It has been the experience of the author that most real-world problems require

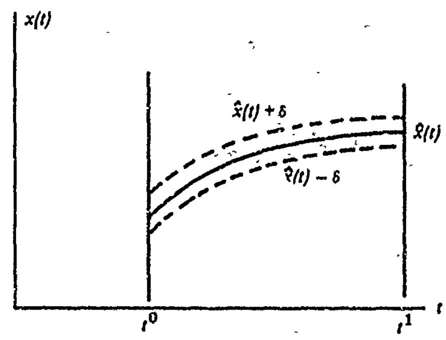


Figure 6-4. A Neighborhood of $\hat{x}(t)$

some modification of the basic optimization problems. Without a thorough knowledge of the theory, the designer will probably have no idea of how to modify the existing theory to suit his purposes.

6-2 THE FUNDAMENTAL PROBLEM OF THE CALCULUS OF VARIATIONS

Examples 6-1 and 6-2 have features in common that allow for the formulation of an entire class of problems containing these two. For the sake of generality, let the variable $x(t)$ be a vector valued function of the real variable t, i.e.,

$$x(t) = \begin{bmatrix} x_1(t) \\ \vdots \\ x_n(t) \end{bmatrix} \qquad (6\text{-}13)$$

where $x_i(t)$ are real valued functions of t

The problem considered here may be formulated as Definition 6-1.

Definition 6-1 (Fundamental Problem of the Calculus of Variations). Find a function $x(t)$ in $C^2(t^0, t^1)$ which satisfies

$x_i(t^0) = x_i^0$, for some indices $1 \le i < n$

$x_j(t^1) = x_j^1$, for some indices $1 < j < n$

(6-14)

and which minimizes

$$J(x) = \int_{t^0}^{t^1} F(t, x, x')dt \qquad (6\text{-}15)$$

where F is a real valued function of all its arguments and

$$x' = \begin{bmatrix} \frac{dx_1}{dt} \\ \vdots \\ \frac{dx_n}{dt} \end{bmatrix} \qquad (6\text{-}16)$$

If the reader wishes he may consider $x(t)$ as being a real valued function of t, the generalization to vector valued functions is simply a matter of notation. The conditions, Eq. 6-14, specify some or all of the components of $x(t)$ at the end points of the interval $t^0 < t < t^1$. This corresponds to demanding that the curves in Examples 6-1 and 6-2 pass through given points.

6-2.1 NECESSARY CONDITIONS FOR THE FUNDAMENTAL PROBLEM

Only necessary conditions for solution of the fundamental problem of Def. 6-1 will be developed here, i.e., the existence of a solution, $\hat{x}(t)$, in $C^2(t^0,t^1)$ first will be assumed. A set of conditions that $\hat{x}(t)$ must satisfy then will be derived. These conditions then may be employed in particular problems to find functions $x(t)$ that are candidate solutions of the problem. Hopefully, there will be just one such candidate that must then be the solution. If there are several candidates, other methods must be used to choose the solution. This problem will be discussed later.

Graphically, the method of obtaining conditions on the solution, $\hat{x}(t)$, of the fundamental problem will be to allow small changes in $\hat{x}(t)$ and examine the behavior of $J(x)$. An admissible, small perturbation is illustrated in Fig. 6-5. The equation for this curve is $\hat{x}(t) + \epsilon\eta(t)$ where ϵ is a small real number and $\eta(t)$ is any member of $C^2(t^0,t^1)$ such that

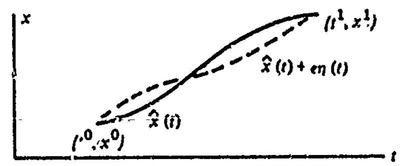


Figure 6-5. Perturbation from Optimum

$$\left.\begin{aligned} \eta_i(t^0) &= 0, \quad \text{for } i \text{ with } x_i(t^0) = x_i^0 \\ \eta_j(t^1) &= 0, \quad \text{for } j \text{ with } x_j(t^1) = x_j^1 \end{aligned}\right\} \quad (6\text{-}17)$$

To examine the effect of this perturbation of $J(x)$, substitute $\hat{x} + \epsilon\eta$ in Eq. 6-15,

$$J(\hat{x} + \epsilon\eta) = \int_{t^0}^{t^1} F(t, \hat{x} + \epsilon\eta, \hat{x}' + \epsilon\eta')dt. \qquad (6\text{-}18)$$

Recall that $\hat{x}(t)$ is a local minimum of $J(x)$ subject to Eq. 6-14, i.e., any small change in $\hat{x}(t)$ increases $J(x)$. For any given function $\eta(t)$ in $C^2(t^0,t')$ and satisfying Eq. 6-17, $\hat{x}(t) + \epsilon\eta(t)$ is in $C^2(t^0,t^1)$ and satisfies Eq. 6-14 for all ϵ. Therefore, for this $\eta(t)$, $J(\hat{x} + \epsilon\eta)$ is a real valued function of ϵ. Further, for $\epsilon = 0$, $J(\hat{x} + \epsilon\eta)$ has a relative minimum and it is assumed that $F(t;x;x')$ is twice continuously differentiable in x and x' so that $J(\hat{x} + \epsilon\eta)$ is a twice continuously differentiable function of ϵ. Theorem 2-2 then applies, and it is required that

$$\frac{\partial}{\partial\epsilon} J(\hat{x} + \epsilon\eta)\Big|_{\epsilon=0} = 0 . \tag{6-19}$$

The object now is to transform condition, Eq. 6-19, into conditions on $\hat{x}(t)$. Performing the differentiation indicated in Eq. 6-19,

$$\frac{\partial}{\partial\epsilon} J(\hat{x} + \epsilon\eta)\Big|_{\epsilon=0} =$$

$$\int_{t^0}^{t^1} \left(\frac{\partial F}{\partial x}\eta + \frac{F}{x'}\eta'\right) dt = 0, \tag{6-20}$$

where the arguments in the partial derivatives of f in Eq. 6-20 are $\hat{x}(t)$ and $\hat{x}'(t)$. It is important to remember that Eq. 6-20 is required to hold for any $\eta(t)$ in $C^2(t^0,t^1)$. which satisfies Eq. 6-17.

Integrating the second term in the integrand of Eq. 6-20 yields

$$\int_{t^0}^{t^1} \left(\frac{\partial F}{\partial x} - \frac{d}{dt}\frac{\partial F}{\partial x'}\right) \eta dt$$

$$+\frac{\partial F}{\partial x'}\left[t^1, \hat{x}(t^1), \hat{x}'(t^1)\right]\eta(t^1) \tag{6-21}$$

$$-\frac{\partial F}{\partial x'}\left[t^0, \hat{x}(t^0), \hat{x}'(t^0)\right]\eta(t^0) = 0.$$

Since the behavior of η inside the interval $t^0 < t < t^1$ and at its en re independent, the integral and boundary ms in Eq. 6-21 may be treated independently, i.e., each is required to be zero. One of the major results which follows is a direct application of Lemma 6-1.

Lemma 6-1: If $M(t)$ is a continuous function on $t^0 < t < t^1$ and if

$$\int_{t^0}^{t^1} M(t)\eta(t)\,dt = 0 \tag{6-22}$$

for all $\eta(t)$ in $C^2(t^0, t^1)$ with $\eta(t^0) = \eta(t^1) = 0$, then $M(t) = 0$, $t^0 < t < t^1$.

The ideas involved in the proof are easily seen graphically. In Fig. 6-6 a point t^*, $t^0 <$

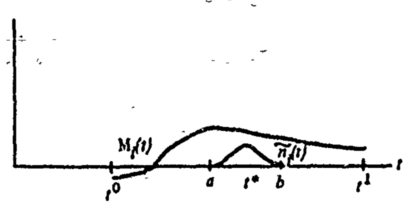


Figure 6-6. Graphical Proof of Lemma 6-1

$t^* < t^1$, is shown where $M_l(t^*) \neq 0$. The curve $\bar{\eta}_l(t)$ is then constructed so that neither function is zero in the interval $a < t < b$. Their integral over the entire interval is then nonzero which is a contradiction of Eq. 6-22, so $M_l(t^*) = 0$.

Since the two terms in Eq. 6-21 must each be zero,

$$\int_{t^0}^{t^1} \left[\frac{\partial F}{\partial x} - \frac{d}{dt}\left(\frac{\partial F}{\partial x'}\right)\right] \eta\, dt = 0 \tag{6-23}$$

for all $\eta(t)$ in $C^2(t^0,t^1)$. In any subinterval of $t^0 < t < t^1$ where $\hat{x}(t)$ is continuously differentiable, the quantity

$[\partial F/\partial x - d/dt\ (\partial F/\partial x')]$ is continuous. Therefore Lemma 6-1 implies

$$\frac{\partial F}{\partial x} - \frac{d}{dt}\left(\frac{\partial F}{\partial x'}\right) = 0 \tag{6-24}$$

in that subinterval.

If, however, $\hat{x}'(t)$ has a jump discontinuity at some point $\bar{t}$ then $[\partial F/\partial x - d/dt\ (\partial F/\partial x')]$ need not be continuous at $\bar{t}$ and Lemma 6-1 may not be applied over any subinterval containing $\bar{t}$. Since Eq. 6-24 must hold in subintervals on both sides of $\bar{t}$, this equation may be integrated from $\bar{t} - \delta$, $\delta > 0$, to t to obtain

$$\frac{\partial F}{\partial x'} = \int_{\bar{t}-\delta}^{t} \frac{\partial F}{\partial x}\, dt + C. \tag{6-25}$$

The vector $\partial F/\partial x$ is piecewise continuous so the right-hand side of Eq. 6-25 is continuous. Therefore $\partial F/\partial x'$ is continuous even at $\bar{t}$.

These results may be stated in the form of a theorem.

Theorem 6-1: The following conditions must be satisfied by the solution of the problem of Def. 6-1, $\hat{x}(t)$, whose derivative is piecewise continuous;

$$\frac{\partial F}{\partial x}[t, \hat{x}(t), \hat{x}'(t)] - \frac{d}{dt}\left\{\frac{\partial F}{\partial x'}[t, \hat{x}(t), \hat{x}'(t)]\right\} = 0 \tag{6-26}$$

at points of continuity of $\hat{x}(t)$

$$\frac{\partial F}{\partial x'}[t^1, x(t^1), \hat{x}'(t^1)]\,\eta(t^1) - \frac{\partial F}{\partial x'}[t^0, \hat{x}(t^0), \hat{x}'(t^0)]\,\eta(t^0) = 0 \tag{6-27}$$

for all $\eta(t^0), \eta(t^1)$ satisfying Eq. 6-14, and

$$\frac{\partial F}{\partial x'}[\bar{t} - 0, \hat{x}(\bar{t} - 0), \hat{x}'(\bar{t} - 0)] = \frac{\partial F}{\partial x'}[\bar{t} + 0, \hat{x}(\bar{t} + 0), \hat{x}'(\bar{t} + 0)] \tag{6-28}$$

at each point $\bar{t}$ of discontinuity of $\hat{x}'(t)$.

Condition, Eq. 6-26, is a second-order differential equation in $\hat{x}(t)$ and is called the Euler-Lagrange equation. Condition, Eq. 6-27, is called a transversality condition. For each i or j such that $\eta_i(t^0)$ or $\eta_j(t^1)$ is not specified by Eq. 6-14, Eq. 6-27 implies $\partial F/\partial x_i'(t^0) = 0$ or $\partial F/\partial x_j'(t^1) = 0$. The condition, Eq. 6-28, at discontinuities (called corners) in $\hat{x}'(t)$ is called the Weierstrass-Erdmann corner condition.

One further necessary condition will be important for further development. Define the Weierstrass E-function as

$$E(t, x, x', w) = F(t, x, w) - F(t, x, x') - \frac{\partial F}{\partial x'}(t, x, x')(w - x'). \tag{6-29}$$

The proof of the Weierstrass necessary condition may be found in Ref. 2, page 149. The result only will be given here as Theorem 6-2.

Theorem 6-2: If the function $\hat{x}(t)$ is the solution of the problem of Def. 6-1, then it is necessary that

$$E[t, \hat{x}(t), \hat{x}'(t), w] \geq 0 \tag{6-30}$$

for all $t^0 < t < t^1$ and all finite w.

The Weierstrass condition of Theorem 6-2 generally is not used to generate candidate solutions of the fundamental problem.

Rather, when solutions of Eq. 6-26 are determined, Eq. 6-30 is used to eliminate unsuitable functions, i.e., it may very well disqualify a function which satisfies Eq. 6-26.

The derivation of necessary conditions for the fundamental problem is only very lightly covered here. Further, the theory of sufficient conditions is completely neglected. For outstanding and complete treatments of these topics see Refs. 2, 3, 4, and 5.

6-2.2 SPECIAL CASES AND EXAMPLES

In many problems the form of the function $F(t,x,x')$ allows for simplification for the Euler-Lagrange equation, Eq. 6-26. In any case, Eq. 6-26 may be written, using the chain rule of differentiation and the notation

$$F_x = \frac{\partial F}{\partial x}, F_{x'} = \frac{\partial F}{\partial x'}, F_{tx'} = \frac{\partial^2 F}{\partial t \partial x'}, \text{ and}$$

$$F_{x'x} = \frac{\partial^2 F}{\partial x' \partial x}$$

to obtain

$$F_x - F_{x't} - x'^T F^T_{x'x} - x''^T F_{x'x'}$$

$$= 0. \tag{6-31}$$

This is simply a second-order differential equation for $x(t)$.

Several special cases with examples will now be considered.

Case 1. F does not depend on x':

$$F = F(t,x). \tag{6-32}$$

Eq. 6-31 in this case is

$$F_x(t,x) = 0. \tag{6-33}$$

This is simply an algebraic equation between t and x. Since there will be no constants of integration, it will not generally be possible to pass the resulting curve through particular points. This means that a solution to such a problem generally will not exist.

Example 6-3: Minimize

$$\int_0^1 x^2 dt$$

for

$$x(0) = 0, \; x(1) = 1.$$

The condition, Eq. 6-33, is

$$2x = 0.$$

But it is, therefore, impossible to satisfy $x(1) = 1$ so the problem has no solution.

To get an idea of what has gone wrong, note that since $x^2(t) > 0$ for each t,

$$\int_0^1 x^2(t)dt > 0$$

for any curve on $0 < t < 1$. It is, therefore, clear that if there were a curve which minimized $\int_0^1 x^2 dt$, then the minimum value of the integral would be non-negative.

It was noted that no minimum exists. However, consider the family of curves

$$x_n(t) = t^n.$$

These curves all satisfy the end conditions and

$$J(x_n) = \int_0^1 t^{2n} dt = \frac{1}{2n+1}.$$

Therefore, it is possible to choose n large enough so that $\int_0^1 x_n^2 dt$ is as close as desired to zero. However, the limit of $x_n(t)$ as n approaches infinity is the function

$$x_\infty(t) = \begin{cases} 0, & t < 1 \\ 1, & t = 1, \end{cases}$$

and this is not even a continuous function. The class of functions $x_n(t)$ are illustrated in Fig. 6-7.

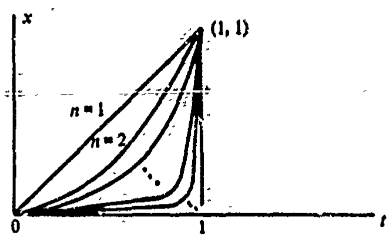


Figure 6-7. Minimizing Sequence

In this illustration, a solution of the problem exists in the class of piecewise continuous functions but not in the class of twice continuously differentiable functions. This problem, therefore, should serve as a warning that not all innocent looking calculus of variations problems have solutions.

Case 2. F depends only on x':

$$F = F(x'). \tag{6-34}$$

Eq. 6-31 is in this case

$$F_{x'x'} x'' = 0. \tag{6-35}$$

Example 6-4: Using the formulation of Example 6-1, find the shortest curve in the t-x plane which passes through the points (0,0) and (1,1).

The function F from Eq. 6-1 is

$$F = [1 + (x')^2]^{1/2}.$$

The form of the Euler-Lagrange equation in Eq. 6-35 applies in this case to yield

$$-[1 + (x')^2]^{-3/2} x'' = 0.$$

Since $(x')^2 \geq 0$, $[1 + (x')^2] \neq 0$ and $x'(t)$ is required to be continuous so $[1 + (x')^2] \neq \infty$ and it, therefore, is required that

$$x''(t) = 0$$

or

$$x(t) = at + b,$$

where a and b are constants. This implies that the shortest path between two points in a plane is a straight line. This shouldn't shake anyone up.

The end conditions yield

$$x(0) = b = 0$$

and

$$x(1) = a = 1.$$

Therefore the solution of the problem is

$$x(t) = t.$$

Case 3. F depends only on t and x':

$$F = F(t, x'). \tag{6-36}$$

Eq. 6-26 is, in this case,

$$\frac{d}{dt} F_{x'}(t, x') = 0$$

or

$$F_{x'}(t,x') = C \tag{6-37}$$

where C is an arbitrary constant.

Case 4. x is a real-valued function, and F depends only on x and x':

$$F = F(x,x'). \tag{6-38}$$

Eq. 6-31 is, in this case,

$$F_x - F_{x'x}\,x' - F_{x'x'}\,x'' = 0.$$

Multiplying by x' yields

$$x'F_x - (x')^2 F_{x'x} - x'x''F_{x'x'} = 0.$$

This is just

$$\frac{d}{dt}(F - x'F_{x'}) = 0,$$

so

$$F - x'F_{x'} = C \tag{6-39}$$

where C is an arbitrary constant.

Example 6-5. Solve the Brachistochrone problem of Example 6-2.

The function F from Eq. 6-4 is

$$F = \left[\frac{1 + (x')^2}{2gx}\right]^{1/2}.$$

Eq. 6-39 applies in this case and yields

$$\left[\frac{1 + (x')^2}{2gx}\right]^{1/2} - \frac{(x')^2}{(2gx)^{1/2}\,[1 + (x')^2]^{1/2}}$$

$$= C.$$

This reduces to

$$1 = \left\{2gx\,[1 + (x')^2]\right\}^{1/2} C,$$

or

$$x[1 + (x')^2] = C_1$$

where C_1 is a new constant.

The solution of this differential equation is a family of cycloids in parametric form

$$t = C_2 + \frac{C_1}{2}(s - \sin s)$$

and

$$x = \frac{C_2}{2}(1 - \cos s).$$

The constants C_1 and C_2 are to be determined so that the cycloid which passes through the given points is fixed.

It should be noted that each of the problems treated here reduced to the solution of a nonlinear differential equation. This is characteristic of problems of the calculus of variations. The reader is undoubtedly aware that it is only in the simplest cases that closed form solutions of these differential equations may be obtained. Further, questions of existence and uniqueness of solutions are by no means trivial.

6-2.3 VARIATIONAL NOTATION AND SECOND-ORDER CONDITIONS

For $J(x) = \int_{t^0}^{t^1} F(t,x,x')dt,$

define the first variation of $J(x)$ as

$$\delta J(x) \equiv \frac{d}{d\epsilon} J(x + \epsilon\delta x)\Big|_{\epsilon=0} \tag{6-40}$$

$$= \frac{d}{d\epsilon} \int_{t^0}^{t^1} F(t, x + \epsilon\delta x, x' + \epsilon\delta x') dt \Big|_{\epsilon = 0}$$

$$= \int_{t^0}^{t^1} \left[\frac{\partial F}{\partial x}(t,x,x')\delta x + \frac{\partial F}{\partial x'}(t,x,x')\delta x' \right] dt.$$

Note that all this does not require that $x(t)$ be the solution of the fundamental problem. If, however, $x(t) = \hat{x}(t)$ is the solution of the fundamental problem, then it is clear from Eq. 6-19 that it is necessary that

$$\delta J(\hat{x}) = 0, \tag{6-41}$$

for all $\delta x(t)$ for which $\hat{x}(t) + \epsilon\delta x(t)$ satisfy the end conditions in the fundamental problem.

In a way quite similar to the definition of the first variation, the second variation may be defined as

$$\delta^2 J(x) = \frac{d^2}{d\epsilon^2} J(x + \epsilon\delta x) \Big|_{\epsilon = 0}$$

Performing the differentiation, this is

$$\delta^2 J(x) = \frac{d^2}{d\epsilon^2} \int_{t^0}^{t^1} F(t, x + \epsilon\delta x, x' + \epsilon\delta x')\, dt \Big|_{\epsilon = 0}$$

$$= \frac{d}{d\epsilon} \int_{t^0}^{t^1} \left[\delta x^T \frac{\partial F^T}{\partial x}(t, x + \epsilon\delta x, x' + \epsilon\delta x') + \delta x'^T \frac{\partial F^T}{\partial x'}(t, x + \epsilon\delta x, x' + \epsilon\delta x') \right] dt \Big|_{\epsilon = 0}$$

$$= \int_{t^0}^{t^1} \left[\delta x^T \frac{\partial^2 F}{\partial x^2}(t,x,x')\delta x^T + 2\delta x^T \frac{\partial^2 F}{\partial x \partial x'}(t,x,x')\delta x' + \delta x'^T \frac{\partial^2 F}{\partial x'^2}(t,x,x')\delta x' \right] dt.$$

Define

$$A = \frac{\partial^2 F}{\partial x^2}$$

$$B = 2\left(\frac{\partial^2 F}{\partial x \partial x'}\right)$$

and

$$C = \frac{\partial^2 F}{\partial x'^2}.$$

With this notation,

$$\delta^2 J(x) = \int_{t^0}^{t^1} (\delta x^T A \delta x + \delta x^T B \delta x' + \delta x'^T C \delta x')\, dt.$$

If $F(t,x,x')$ has three derivatives, then by Taylor's formula

$$J(x + \epsilon\delta x) = J(x) + \frac{dJ}{d\epsilon}\Big|_{\epsilon = 0} \epsilon + \frac{d^2 J}{d\epsilon^2}\Big|_{\epsilon = 0} \epsilon^2 + \frac{d^3 J}{d\epsilon^3}\Big|_{\epsilon = \bar{\epsilon}} \epsilon^3 \tag{6-42}$$

where $0 < \bar{\epsilon} < \epsilon$. If we computed $d^3J/d\epsilon^3$, it would involve a sum of terms each containing third degree terms in δx and $\delta x'$. For δx and $\delta x'$ sufficiently small, this term may be neglected to obtain a second-order approximation when $\epsilon = 1$ so that

$$J(\hat{x} + \delta x) \approx J(x) + \delta J + \delta^2 J.$$

It is clear then that δJ and $\delta^2 J$ play the role of differentials in the theory of functionals.

Further, if $\hat{x}(t)$ yields a relative minimum for the fundamental problem, then $J(\hat{x} + \epsilon\delta x)$ is a relative minimum at $\epsilon = 0$. It is, therefore, necessary that

$$\left.\frac{d^2 J}{d\epsilon^2}\right|_{\epsilon=0} \geq 0.$$

This is just

$$\delta^2 J(\hat{x}) \geq 0,$$

or

$$\int_{t^0}^{t^1} (\delta x^T A \delta x + \delta x^T B \delta x' + \delta x'^T C \delta x')\,dt \geq 0 \qquad (6\text{-}43)$$

for all $\delta x(t)$ such that $\hat{x} + \delta x$ satisfy the end conditions for the fundamental problem. In what follows it will be convenient to limit $\delta x(t)$ to those variations which satisfy $\delta x(t^0) = \delta x(t^1) = 0$.

If $\delta x'(t)$ is small for all t, then $\delta x(t)$ must also be small since $\delta x(t^0) = 0$. On the other hand, it is possible to choose $\delta x(t)$ which is zero at the endpoints and small for all t, but which has large derivatives. One might, therefore, be led to believe that the derivative term in the inequality of Eq. 6-43 is dominant. This would then require C to be positive semi-definite.

To show that this is the case, assume that there is a point t^*, $t^0 < t < t^1$ and a nonzero vector h such that $h^T C(t^*) h = -2\beta < 0$. For any continuous $\delta x'(t)$ such that $\delta x'(t^*) = h$, there is an interval

$t^* - \alpha < t < t^* + \alpha > 0$, such that

$$\delta x'(t)^T C(t) \delta x'(t) < -\beta < 0$$

in $t^* - \alpha < t < t^* + \alpha$.

Define

$$\delta x(t) = \begin{cases} \dfrac{\alpha}{\pi} h \sin\left[\dfrac{\pi(t - t^*)}{\alpha}\right], & t^* - \alpha < t < t^* + \alpha \\ 0, & \text{elsewhere} \end{cases}$$

so that

$$\delta x'(t) = \begin{cases} h \cos\left[\dfrac{\pi(t - t^*)}{\alpha}\right], & t^* - \alpha < t < t + \alpha \\ 0, & \text{elsewhere} \end{cases}$$

Now, Eq. 6-43 is

$$0 < \int_{t^0}^{t^1} (\delta x^T A \delta x + \delta x^T B \delta x' + \delta x'^T C \delta x')\,dt$$

$$= \int_{t^* - \alpha}^{t^* + \alpha} \left\{ \frac{\alpha^2}{\pi^2} \sin^2\left[\frac{\pi(t - t^*)}{\alpha}\right] h^T A h \right.$$

$$+\frac{\alpha}{\pi}\sin\left[\frac{\pi(t-t^*)}{\alpha}\right]\cos\left[\frac{\pi(t-t^*)}{\alpha}\right]h^TBh$$

$$\left.+\cos^2\left[\frac{\pi(t-t^*)}{\alpha}\right]h^TCh\right\}dt$$

$$<\frac{\alpha^3}{\pi^3}M\int_{-\pi}^{\pi}\sin^2\theta d\theta$$

$$+\frac{\alpha^2}{\pi^2}P\int_{-\pi}^{\pi}|\sin\theta||\cos\theta|d\theta$$

$$-\beta\frac{\alpha}{\pi}\int_{-\pi}^{\pi}\cos^2\theta d\theta,$$

where

$$M=\max_t |h^TA(t)h|$$

$$P=\max_t |h^TB(t)h|$$

Therefore, integration in the preceding inequality yields

$$0<\alpha^2\left(\frac{\alpha}{\pi^2}M+\frac{2}{\pi^2}P-\frac{1}{\alpha}\beta\right)$$

However, since α may be chosen arbitrarily small and $\beta > 0$ the right side will be negative for sufficiently small α. But this is a contradiction. Therefore, the assumption that there exist t^* and h such that $h^TCh < 0$ is incorrect. This implies $h^TCh \geq 0$ for all h and t^*. Therefore $C(t)$ is positive semi-definite. Since

$$C(t)=\frac{\partial^2 F}{\partial x'^2}$$

this result may be stated as Theorem 6-3.

Theorem 6-3: A necessary condition for the fundamental problem to have a relative minimum at $\hat{x}(t)$ is that

$$\frac{\partial^2 F}{\partial x'^2}[t,\hat{x}(t),\hat{x}'(t)]$$

be positive semi-definite for all t, $t^0 \leq t \leq t^1$.

Gelfand and Fomin (Ref. 2, p. 104) indicate that people are prone to argue that positive definiteness of $\partial^2F/\partial x'^2$ at each point of the solution is a sufficient condition for an extremum. They point out, however, that this is not the case and, in fact, that no local condition can provide sufficient conditions. For a treatment of sufficient conditions see Refs. 2, 3, 4, and 5.

6-2.4 DIRECT METHODS

The direct methods of the calculus of variations seek to generate a sequence of functions $\{x^{(n)}(t)\}$ such that, if ξ is the infimum of $J(x)$ over all admissible x, then

$$\lim_{n\to\infty} J[x^{(n)}]=\xi. \tag{6-44}$$

Direct methods are capable of showing existence of solution as well as construction of approximations of the solution. It is generally very difficult to prove existence of a solution of the nonlinear boundary-value problem in the necessary conditions for the fundamental problem. It is often possible, however, to show that the sequence $[x^{(n)}(t)]$ converges to a function $\hat{x}(t)$ which is the solution of the fundamental problem, i.e.,

$$\lim_{n \to \infty} J[x^{(n)}] = J(\hat{x}) = \xi . \qquad (6\text{-}45)$$

It is clear, however, that a sequence which satisfies Eq. 6-44 may very well fail to converge to an admissible function $\hat{x}(t)$. This must necessarily be the case if no solution of the fundamental problem exists. From an engineering point of view, one may not be too concerned with existence of a limit of the sequence $[x^{(n)}(t)]$. Provided it is possible to successively reduce J, consistently better results are being obtained and the process will be continued until no further meaningful reduction in J may be achieved. For an outstanding treatment of convergence of direct methods, see Ref. 2, page 192, and Ref. 3, page 127.

The problem of primary interest to the engineer is the construction of a minimizing sequence. There are many ways of generating such a sequence, only two of which will be treated here. These methods are known as the Ritz Method and the Method of Finite Differences.

6-2.4.1 THE RITZ METHOD

The Ritz Method is based on the idea of representing functions by using linear combinations of known functions; i.e., given $\phi_i(t)$, $i = 1, 2, \ldots$, which preferably form a complete set, a function is represented by

$$y(t) = \sum_{i=1}^{\infty} a_i \phi_i(t)$$

where a_i are constants. Classical, trigonometric Fourier series is an example of this kind of representation.

In the Ritz Method, the nth function in the minimizing sequence is formed by

$$x^{(n)}(t) = \sum_{i=1}^{n} a_i \phi_i(t) \qquad (6\text{-}46)$$

where the $\phi_i(t)$ are chosen so that $x^{(n)}(t)$ satisfies the end conditions associated with the fundamental problem. This expression is then substituted into $J(x)$ to obtain

$$J[x^{(n)}] =$$

$$\int_{t^0}^{t^b} F\left[t, \sum_{i=1}^{n} a_i \phi_i(t), \sum_{i=1}^{n} a_i \phi_i'(t)\right] dt. \qquad (6\text{-}47)$$

The object now is to choose the coefficients a_i, $i = 1, \ldots, n$, so that $J[x^{(n)}]$ is as small as possible. For this purpose, it should be noted that the right side of Eq. 6-47 is simply a function of n parameters. The problem is now to minimize this function without any other constraints. For this purpose, any of the methods of Chapter 2 may be used.

The property

$$J[x^{(n+1)}] < J[x^{(n)}]$$

follows readily from the method of determining the a_i. It is clear that by choosing $a_{n+1} = 0$, $x^{(n+1)}(t) = x^{(n)}(t)$. However, by allowing a_{n+1} to be nonzero, a larger number of functions are available as candidates for minimum of $J[x^{(n+1)}]$ than $J[x^{(n)}]$. The minimum of $J[x^{(n+1)}]$ will, therefore, certainly not be greater than that of $J[x^{(n)}]$ and this is the desired result.

In practice, the rate of convergence of $[x^{(n)}(t)]$ depends strongly on the functions $\phi_i(t)$ chosen. The number of terms required to obtain a reasonable approximation of the solution is greatly reduced if these functions are chosen based on a reasonable engineering estimate of the form of the solution. By making a judicious choice of the $\phi_i(t)$, a good approximation of the solution may be obtained with as few as two or three terms.

Example 6-6: In solving the boundary-value problem

$$x'' + (1 + t^2)x + 1 = 0$$

$$x(-1) = x(1) = 0,$$

it is necessary to minimize the functional

$$J(x) = \int_{-1}^{1} [x'^2 - (1 + t^2)x^2 - 2x]\, dt \tag{6-48}$$

subject to the end conditions $x(-1) = x(1) = 0$.

In order to minimize $J(x)$ of Eq. 6-48, by the Ritz Method, choose

$$\phi_i(t) = (1 - t^{2i}).$$

If for a first approximation $n = 2$ is chosen,

$$x^{(2)}(t) = a_1(1 - x^2) + a_2(1 - x^4). \tag{6-49}$$

Substituting $x^{(2)}$ into Eq. 6-48 and integrating yields

$$J(x^{(2)}) = 8\left(\frac{19}{105}a_1^2 + \frac{10}{45}2a_1a_2 + \frac{1244}{3465}a_2^2 - \frac{1}{3}a_1 - \frac{2}{5}a_2\right).$$

This is a positive definite quadratic form, so it has a unique minimum which may be obtained by setting its first derivatives equal to zero. This yields

$$\frac{38}{105}a_1 + \frac{20}{45}a_2 - \frac{1}{3} = 0$$

$$\frac{20}{45}a_1 + \frac{2488}{3465}a_2 - \frac{2}{5} = 0.$$

The solution of these equations is

$$a_1 = 0.9877$$

$$a_2 = -0.05433.$$

Substituting these coefficients into Eq. 6-49,

$$x^{(2)}(t) = \frac{1}{4252}(3969 - 4200t^2 + 231t^4).$$

In particular,

$$x^{(2)}(0) = 0.93344.$$

If the three term approximation

$$x^{(3)} = a_1(1 - t^2) + a_2(1 - t^4) + a_3(1 - t^6)$$

is determined in the same manner,

$$x^{(3)}(0) = 0.93207.$$

This might lead one to believe that both $x^{(2)}(t)$ and $x^{(3)}(t)$ are good approximations of the solution.

6-2.4.2 METHOD OF FINITE DIFFERENCES

The Method of Finite Differences, as its name implies, is simply based on the replacement of derivatives and the integral by finite approximations of these continuous operations. A grid, $t^0 = t_0, \; t_1, \ldots, t_{n+1} = t^1$, is placed on the interval $t^0 < t < t^1$ and the value of $x(t)$ only at the grid points is sought, i.e., only the parameters $x_i = x(t_i)$, $i = 1, \ldots, n$, and perhaps $x(t_0)$ or $x(t_{n+1})$, are sought.

Replacing derivatives by finite differences and the integral by a finite sum, the problem is to determine the x_i which minimize

$$J(x_i) = \sum_{i=0}^{n} \left\{ F\left[t_i, x_i, \frac{x_{i+1} - x_i}{t_{i+1} - t_i}\right] \times (t_{i+1} - t_i) \right\}.$$

The problem is now simply an unconstrained minimization problem in a finite dimensional space and may be solved by the methods of Chapter 2.

6-3 A PROBLEM OF BOLZA

6-3.1 STATEMENT OF THE PROBLEM

Many real-world optimal design problems cannot realistically be reduced to the finite dimensional form of Chapter 5. In many problems the system varies continuously in time or space, so functions rather than just parameters must be determined. Examples 6-1 and 6-2, par. 6-1, are extremely simple, yet even they involve distribution of the controlling factor over space and time

As has been seen in previous chapters, optimal design problems involve ideas of design variables and state variables. Further, since the system being designed must be capable of performing certain functions, side conditions on the state and design variables occur. It has been observed in previous chapters that these side conditions generally include both equality and inequality constraints. An extension to inequality constraints will be given in par. 6-4.

The problem to be treated here is given in Definition 6-2.

Definition 6-2 (Problem of Bolza): The problem of Bolza is a problem of finding $u(t)$, b, $x(t)$, $t^0 < t < t^n$, which minimizes

$$J = g_0(b, t^j, x^j) + \int_{t^0}^{t^n} f_0\,[t, x(t), u(t), b]\,dt \tag{6-50}$$

subject to the conditions

$$\frac{dx}{dt} = f(t, x, u, b), \; t^0 < t < t^n \tag{6-51}$$

$$g_\alpha(b, t^j, x^j) + \int_{t^0}^{t^n} L_\alpha\,[t, x(t), u(t), b]\,dt = 0, \; \alpha = 1, \ldots, r \tag{6-52}$$

$$\phi_\beta(t, x, u, b) = 0, \quad \beta = 1, \ldots, q, \; t^0 < t < t^n \tag{6-53}$$

where

$$x(t) = \begin{bmatrix} x_1(t) \\ \vdots \\ x_n(t) \end{bmatrix}, \quad u(t) = \begin{bmatrix} u_1(t) \\ \vdots \\ u_m(t) \end{bmatrix},$$

$$b = \begin{bmatrix} b_1 \\ \vdots \\ b_\ell \end{bmatrix} \tag{6-54}$$

$$f(t,x,u,b) = \begin{bmatrix} f_1(t,x,u,b) \\ \cdot \\ \cdot \\ \cdot \\ f_n(t,x,u,b) \end{bmatrix},$$

and $t^0 < t^j < t^\eta$, where (t^j,x^j) are intermediate points, $j = 1, ..., \eta - 1$.

For the problem considered here it will be required that the conditions, Eqs. 6-53, shall not determine any component of $x(t)$ explicitly. This is equivalent to requiring that the rank of the matrix

$$\left[\frac{\partial \phi_\beta}{\partial u_k}(t,x,u,b)\right]_{q \times m} \qquad (6\text{-}55)$$

shall be q for all admissible values of the arguments. In case some constraint function should depend only on $x(t)$ and t, this constraint is called a state variable constraint. This kind of constraint will be discussed in a later paragraph.

The vector variable $x(t)$ is called the state variable, $u(t)$ is called the design (or control) variable, and b is called the design (or control) parameter. Eqs. 6-52 contain the boundary conditions on the state variable and functions which determine the end points of the interval, t^0 and t^η. The independent variable t may be time or a space-type variable, depending on the problem being considered.

The functions f_0, f, L_α, and ϕ_α are assumed to be continuously differentiable at all points except (t^j,x^j), $j = 1, ..., \eta - 1$. At these points the functions may have jump discontinuities; i.e., the func will have limits along any path, but limits along different paths may have different values. In general, even for systems with very regular functions, $u(t)$ may have jump discontinuities. Therefore $u(t)$ is expected to be only piecewise continuous. The resulting state $x(t)$, therefore, will have only a piecewise continuous derivative in general

The allowed discontinuities of f_0, f, L_α, and ϕ_α play an important role in many real-world problems. This feature allows for completely different forms of state equations, constraints, etc., for different ranges of state and time. It is, therefore, possible to routinely account for sudden changes in system behavior such as reverse in direction of frictional force, motion of objects in a space where physical barriers or restraint surfaces exist, logic built into the system which changes configuration as in staging of rockets, etc. It should be clear that these features are required in order to treat many realistic problems.

For a discussion of the effect of these discontinuities or more detailed necessary conditions and sufficient conditions see Ref. 8.

6-3.2 A MULTIPLIER RULE

As mentioned in par. 6-3.1, real-world optimal design problems require at least the complexity of the Bolza problem of Def. 6-2. In fact, the system designer requires all the tools the mathematical theory of optimal processes can give him. This requirement points out one of the obstacles to engineers in utilizing the modern theories of mathematics. This text cannot possibly present the mathematical theory required of the research mathematician who is developing the theory of optimization. The approach taken here to by-pass this obstacle is to accept a key theorem of Functional Analysis and then proceed to develop the tools required for solving problems of optimal design. A very powerful theorem of Liusternik and Sobolev, Ref. 6, page 209, will be used to obtain necessary conditions for the problem of Bolza.

Theorem 6-4: If $\hat{u}(t)$, $\hat{b}$, t^j, and $\hat{x}(t)$ provide a solution to the Bolza problem of Def. 6-2, then there exist multipliers $\lambda_0 \geq 0$, γ_α, $\alpha = 1, ..., r$, $\lambda_i(t)$, $i = 1, ..., n$, and $\mu_\beta(t)$, $\beta = 1, ..., q$, not all zero, such that

$$\delta \bar{J} = 0, \tag{6-56}$$

where

$$\begin{aligned} \bar{J} = {} & \lambda_0 g_0(b,t^j,x^j) + \sum_{\alpha=1}^{r} \gamma_\alpha g_\alpha(b,t^j,x^j) \\ & + \int_{t^0}^{t^1} \Bigg\{ \lambda_0 f_0(t,x,u,b) \\ & + \sum_{i=1}^{n} \lambda_i(t) \left[\frac{dx_i}{dt} - f_i(t,x,u,b) \right] \\ & + \sum_{\alpha=1}^{r} \gamma_\alpha L_\alpha(t,x,u,b) \\ & + \sum_{\beta=1}^{q} \mu_\beta(t) \phi_\beta(t,x,u,b) \Bigg\} \, dt. \end{aligned} \tag{6-57}$$

Note that the symbol $\delta\bar{J}$ is the first variation of $\bar{J}$ as defined in par. 6-2. For proofs of this multiplier rule, the reader is referred to the literature (Refs. 2,5-9).

This theorem says nothing about the continuity and differentiability properties of the solution $\hat{x}(t)$, $\hat{u}(t)$, and the multipliers $\lambda_i(t)$ and $\mu_\beta(t)$. In general, piecewise continuity is all that may be expected of $u(t)$. Eq. 6-51 then implies $x(t)$ has a piecewise continuous derivative. The properties of $\lambda_i(t)$ and $\mu_\beta(t)$ will be determined when necessary conditions are derived.

6-3.3 NECESSARY CONDITIONS FOR THE BOLZA PROBLEM

The simplest Bolza problem is the one having all its functions three times continuously differentiable. Even in this case, however, $u(t)$ may be only piecewise continuous. To include this possibility, let t^* be a point of discontinuity of any component of $u(t)$.

Before computing the variation called for in Eq. 6-56, it should be noted that t^0; t^j, t^*, and t^n are not fixed but must be determined. This means that these special points must be treated as parameters that are to be determined, much as the design parameter b. At first glance, this may seem to introduce no essential complication into the problem. The behavior of the allowed variations in $x(t)$, however, must be treated very carefully.

Let $\bar{t}$ be a typical point t^j or t^* where $\dot{x}(t)$ may very well be discontinuous. The function $x(t)$ will be changed at $\bar{t}$ by both the independent variation in $x(t)$, $\delta x(t)$, and the shift in the point $(\bar{t}, \delta\bar{t})$. Denote the total change in $x(\bar{t})$ due to both of these sources by $\Delta x(\bar{t})$. It must be assumed that there are no other points t^j or t^* arbitrarily near $\bar{t}$, so limits from the left and right exist. For $t \neq \bar{t}$, $\dot{x}(t)$ is continuous so the total change $\Delta x(t)$ in $x(\bar{t})$ due to $\delta x(t)$ and $\delta\bar{t}$ is continuous and

$$\Delta x(\bar{t}) \equiv \Delta x(\bar{t} - 0) = \Delta x(\bar{t} + 0),$$

where

$$\Delta x(\bar{t} - 0) = \delta x(\bar{t} - 0) + \dot{x}(\bar{t} - 0)\,\delta\bar{t}$$

and

$$\Delta x(\bar{t} + 0) = \delta x(\bar{t} + 0) + \dot{x}(\bar{t} + 0)\,\delta\bar{t}.$$

It should be noted that this condition imposes restrictions on $\delta x(\bar{t} - 0)$ and $\delta x(\bar{t} +$

0). In particular, they are not necessarily the same so $\delta x(t)$ on $t^0 < t < t^n$ need not necessarily be continuous.

Before enforcing Eq. 6-56, put $\bar{J}$ in the form

$$\bar{J} = \lambda_0 g_0(b,t^j,x^j) + \sum_{\alpha=1}^{r} \gamma_\alpha g_\alpha(b,t^j,x^j)$$

$$+ \int_{t^0}^{t^j} \left\{ \lambda_0 f_0(t,x,u,b) + \sum_{i=1}^{n} \lambda_i(t) \right.$$

$$\times \left[\frac{dx_i}{dt} - f_i(t,x,u,b) \right] + \sum_{\alpha=1}^{r} \lambda_\alpha L_\alpha(t,x,u,b)$$

$$\left. + \sum_{\beta=1}^{q} \mu_\beta(t)\phi_\beta(t,x,u,b) \right\} dt$$

$$+ \int_{t^j}^{t} \{\ \} \ dt + \int_{t^*}^{t^n} \{\ \} \ dt \tag{6-58}$$

where the argument of the pair of braces is the same as that of the integrand of the first integral. Note that t^* and t^j are simply typical elements of their respective classes.

For convenience in the development that follows, define

$$G = \lambda_0 g_0(b,t^j,x^j) + \sum_{\alpha=1}^{r} \gamma_\alpha g_\alpha(b,t^j,x^j)$$

$$H(t,x,u,b,\lambda,\gamma,\mu) = \lambda^T(t) f(t,x,u,b)$$

$$- \lambda_0 f_0(t,x,u,b)$$

$$- \sum_{\alpha=1}^{r} \gamma_\alpha L_\alpha(t,x,u,b)$$

$$- \sum_{\beta=1}^{q} \mu_\beta(t)\phi_\beta(t,x,u,b) \tag{6-59}$$

so that Eq. 6-58 becomes

$$\bar{J} = G + \int_{t^0}^{t^j} \left[\lambda^T(t)\frac{dx}{dt} - H \right] dt$$

$$+ \int_{t^j}^{t^*} \left[\lambda^T(t)\frac{dx}{dt} - H \right] dt \tag{6-60}$$

$$+ \int_{t^*}^{t^n} \left[\lambda^T(t)\frac{dx}{dt} - H \right] dt.$$

Eq. 6-56 may now be applied to yield

$$0 = \frac{\partial G}{\partial x^0}\Delta x^0 + \ldots + \frac{\partial G}{\partial x^n}\Delta x^n + \frac{\partial G}{\partial t^0}\delta t^0 + \ldots$$

$$+ \frac{\partial G}{\partial t^n}\delta t^n + \frac{\partial G}{\partial b}\delta b$$

$$+ \int_{t}^{t^j} \left[\lambda^T(t)\frac{d\delta x}{dt} - \frac{\partial H}{\partial x}\delta x \right.$$

$$\left. - \frac{\partial H}{\partial u}\delta u - \frac{\partial H}{\partial b}\delta b \right] dt$$

$$+ \int_{t^j}^{t^*} \left[\lambda^T(t)\frac{d\delta x}{dt} - \frac{\partial H}{\partial x}\delta x \right.$$

$$\left. - \frac{\partial H}{\partial u}\delta u - \frac{\partial H}{\partial b}\delta b \right] dt$$

$$+ \int_{t^*}^{t^n} \left[\lambda^T(t)\frac{d\delta x}{dt} - \frac{\partial H}{\partial x}\delta x \right.$$

$$\left. \frac{\partial H}{\partial u}\delta u \quad \frac{\partial H}{\partial b}\delta b \right] dt$$

$$+\left[\lambda^T(t)\frac{dx(t)}{dt}-H(t)\right]\Bigg|_{t^j+0}^{t^j-0}\cdot\delta t^j$$

$$-\left[\lambda^T(t^0+0)\frac{dx(t+0)}{dt}-H(t^0+0,x,u,b)\right]\delta t^0$$

$$+\left[\lambda^T(t)\frac{dx(t)}{dt}-H(t)\right]\Bigg|_{t^*+0}^{t^*-0}\cdot\delta t^*$$

$$+\left[\lambda(t^n-0)\frac{dx(t^n-0)}{dt}-H(t^n-0,x,u,b)\right]\delta t^n.$$

Integrating the first terms in each integrand by parts yields

$$0=\frac{\partial G}{\partial x^0}\Delta x^0+\ldots+\frac{\partial G}{\partial x^n}\Delta x^n+\frac{\partial G}{\partial t^0}\delta t^0$$

$$+\ldots+\frac{\partial G}{\partial t^n}\delta t^n+\frac{\partial G}{\partial b}\delta b$$

$$-\int_{t^0}^{t^j}\left[\frac{d\lambda^T(t)}{dt}\delta x+\frac{\partial H}{\partial x}\delta x+\frac{\partial H}{\partial u}\delta u+\frac{\partial H}{\partial b}\delta b\right]dt$$

$$-\int_{t^j}^{t^*}[\quad]\,dt-\int_{t^*}^{t^n}[\quad]\,dt$$

$$+[\lambda^T(t^j-0)-\lambda^T(t^j+0)]\Delta x^j$$

$$-[H(t^j-0)-H(t^j+0)]\delta t^j$$

$$+[\lambda^T(t^*-0)-\lambda^T(t^*+0)]\Delta x(t^*)$$

$$-[H(t^*+0)-H(t^*+0)]\delta t^*$$

$$+\lambda^T(t^n-0)\Delta x^n-H(t^n-0)\delta t^n$$

$$-\lambda^T(t^0+0)\Delta x^0+H(t^0+0)\delta t^0.$$

Since the variables Δx^j, δt^j, δb, $\delta x(t)$, and $\delta u(t)$ are arbitrary, (Δx^j are taken as arbitrary along with δt^j so that $\delta x^j=\Delta x^j-\dot{x}^j\delta t^j$ is fixed) Lemma 6-1 applies. Application of this Lemma yields

Theorem 6-5: If $(x(t), u(t), b, t^j, x^j)$ is a solution of the Bolza problem of Def. 6-2, then there exist multipliers $\lambda_0>0$, $\lambda_i(t)$, $i=1,\ldots,n$, γ_α, $\alpha=1,\ldots,r$, $\mu_\beta(t)$, $\beta=1,\ldots,q$, not all zero, satisfying the conditions:

$$\frac{d\lambda}{dt}=-\frac{\partial H^T}{\partial x},\ \text{for } t\neq t^j \tag{6-61}$$

$$\frac{\partial H}{\partial u}=0,\ \text{for } t\neq t^j \tag{6-62}$$

$$\frac{\partial G}{\partial b}-\int_{t^0}^{t^n}\frac{\partial H}{\partial b}\,dt=0 \tag{6-63}$$

$$\left.\begin{aligned}&\frac{\partial G^T}{\partial x^0}-\lambda(t^0)=0\\&\frac{\partial G^T}{\partial x^n}+\lambda(t^n)=0\\&\frac{\partial G^T}{\partial x^j}+\lambda(t^j-0)-\lambda(t^j+0)=0\end{aligned}\right\} \tag{6-64}$$

$$\left.\begin{aligned}&\frac{\partial G}{\partial t^0}+H(t^0+0)=0\\&\frac{\partial G}{\partial t^n}-H(t^n-0)=0\\&\frac{\partial G}{\partial t^j}-H(t^j-0)+H(t^j+0)=0\end{aligned}\right\} \tag{6-65}$$

$$H(t^*-0)-H(t^*+0)=0 \tag{6-66}$$

$$\lambda(t^* - 0) - \lambda(t^* + 0) = 0 \qquad (6\text{-}67)$$

Note that the necessary conditions, Eqs. 6-61 through 6-67, are linear and homogeneous in the multipliers $\lambda_0, \lambda_j(t); \gamma_\alpha, \mu_\beta(t)$. It is, therefore, permissible to choose the magnitude of one multiplier so that the remaining multipliers will be uniquely determined. It seems reasonable that if the necessary conditions obtained by setting $\delta \bar{J} = 0$ are to be related to minimization of J, then λ_0 should not be zero. This is indeed the case and if λ_0 is required to be zero by the necessary conditions, then the Bolza problem is "abnormal" in a sense. Most meaningful problems are normal as defined in Refs. 7, 8, and 9 and require $\lambda_0 \neq 0$. In solving problems using the necessary conditions of Theorem 6-5, one should first verify that Eqs. 6-61 through 6-67 have no solution if $\lambda_0 = 0$. It is then permissible to put $\lambda_0 = 1$ so that the remaining multipliers are uniquely determined.

Even though Eqs. 6-61 through 6-67 are very complicated, it is interesting to note that they provide just the right number of equations to solve for all the unknowns. Eqs. 6-51 along with Eq. 6-61 form a system of $2n$ first-order differential equations for $x(t)$ and $\lambda(t)$. Further, the first and last members of Eq. 6-64 may be considered as $2n$ equations in boundary conditions on λ and x. This is the proper number of boundary conditions. The second equation of Eqs. 6-64 provides any jump conditions in $\lambda(t)$ at the intermediate points t^j, $0 < j < \eta$. Eqs. 6-65 may be interpreted as determining t^j, $j = 0, 1, ..., \eta$, and Eq. 6-66 determines t^*. Eq. 6-67 simply states that λ is continuous even at jump discontinuities in u. Finally, Eq 6-63 determines the design parameter b.

It should be clear that this argument only shows that there are the proper number of equations to determine the unknowns. It does not assert that a solution of Eqs. 6-61 through 6-67 exists. Existence theory for these problems is a difficult question that is treated in Refs. 10 and 11.

The conditions of Theorem 6-5 are very nearly the famous Pontryagin Maximum Principle (Ref. 12). The condition that completes the Maximum Principle is an inequality which follows from the Weierstrass condition of the calculus of variations. This condition is given as Theorem 6-6.

Theorem 6-6: In addition to the conditions of Theorem 6-5, the solution of the Bolza problem must satisfy the condition:

$$H[t, x(t), U, b, \lambda(t), \gamma, 0]$$
$$\leq H[t, x(t), u(t), b, \lambda(t), \gamma, 0] \qquad (6\text{-}68)$$

for all admissible U and all t, $t^0 \leq t \leq t^\eta$.

For proof of this theorem see Refs. 8 and 13.

Another useful result is the following identity:

$$\frac{dH}{dt} = \frac{\partial H}{\partial t}, \text{ for } t \neq t^j. \qquad (6\text{-}69)$$

This condition is useful in case H does not depend explicitly on t. Then H is constant between the points t^j, and at these points it may have discontinuities governed by the third equation in Eq. 6-65.

To prove this relation holds, compute formally

$$\frac{dH}{dt} = \frac{\partial H}{\partial t} + \frac{\partial H}{\partial u}\frac{du}{dt} + \frac{\partial H}{\partial x}\frac{dx}{dt} + \frac{\partial H}{\partial \lambda}\frac{d\lambda}{dt}.$$

Since $\frac{dx}{dt} = f$, $\frac{\partial H}{\partial u} = 0$, $\frac{\partial H}{\partial x} = -\frac{d\lambda^T}{dt}$,

and $\frac{\partial H}{\partial \lambda} = f^T$, this is

$$\frac{dH}{dt} = \frac{\partial H}{\partial t} - \frac{d\lambda^T}{dt} f + f^T \frac{d\lambda}{dt} = \frac{\partial H}{\partial t}$$

as required.

6-3.4 APPLICATION OF THE BOLZA PROBLEM

In order to obtain familiarity with the Bolza problem, several examples will be considered. In order to illustrate the basic ideas associated with the Bolza problems, these examples will be elementary. In real-world problems the engineer should be prepared for complexity that will probably force him to use a numerical method of solution. For examples of the Bolza problem in the field of aerodynamics, a field which contributed greatly to optimal design theory, see Refs. 15,16, and 17.

Example 6-6: Maximum Range Rocket-assisted Projectile

A projectile of mass m is acted on by a fixed force F as shown in Fig. 6-8. The angle of $\theta(t)$ is measured from the x-axis, where the

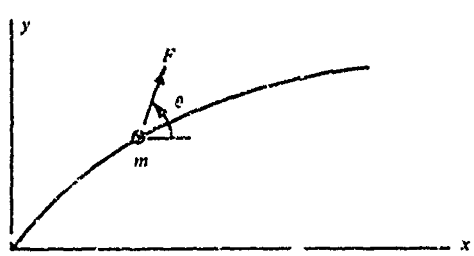


Figure 6-8. Particle in Motion

x- and y-axes are fixed in inertial space. In the problem discussed here, $\theta(t)$ is to be chosen so as to direct the motion of the particle. Hence $\theta(t)$ is the control variable.

Denoting horizontal and vertical components of velocity of the projectile by u and v, respectively, the motion of the projectile is governed by the equations

$$\left.\begin{aligned} \dot{x} &= u \\ \dot{y} &= v \\ \dot{u} &= \frac{F}{m}\cos\theta \\ \dot{v} &= \frac{F}{m}\sin\theta - g, \end{aligned}\right\} \qquad (6\text{-}70)$$

where $\dot{} = \frac{d}{dt}$

The projectile is fired from a gun at time $t = 0$ with $x(0) = y(0) = 0$ and initial velocity $u(0) = V\cos\theta_0$, $v(0) = V\sin\theta_0$, where V is the muzzle velocity of the projectile. The problem at hand is to choose θ_0 and $\theta(t)$ so that at some future time T, the projectile will hit the earth as far as possible from the launch point, i.e., $y(T) = 0$, $x(T)$ = maximum.

In the notation of the Bolza problem, $\theta(t)$ is a design or control variable $u(t)$, θ_0 is a design parameter b, T is terminal time t^1, and (x,y,u,v) is the state. The quantity to be minimized is

$$J = g_0(b, t^1, x^1) = -x(T).$$

Boundary conditions on the state variables are

$$\left.\begin{aligned} g_1 &= x(0) = 0 \\ g_2 &= y(0) = 0 \\ g_3 &= u(0) - V\cos\theta_0 = 0 \\ g_4 &= v(0) - V\sin\theta_0 = 0 \\ g_5 &= y(T) = 0. \end{aligned}\right\} \quad (6\text{-}71)$$

As defined in Eq. 6-59,

$$G = -\lambda_0 x(T) + \gamma_1 x(0) + \gamma_2 y(0) + \gamma_3 [u(0) - V\cos\theta_0] + \gamma_4 [v(0) - V\sin\theta_0] + \gamma_5 y(T) \quad (6\text{-}72)$$

$$H = \lambda_x u + \lambda_y v + \lambda_u \frac{F}{m}\cos\theta + \lambda_v \left(\frac{F}{m}\sin\theta - g\right), \quad (6\text{-}73)$$

where variable named subscripts are used for the λ's.

Theorem 6-5 yields as necessary conditions

$$\left.\begin{aligned} \dot{\lambda}_x &= -\frac{\partial H}{\partial x} = 0 \\ \dot{\lambda}_y &= -\frac{\partial H}{\partial y} = 0 \\ \dot{\lambda}_u &= -\frac{\partial H}{\partial u} = -\lambda_x \\ \dot{\lambda}_v &= -\frac{\partial H}{\partial v} = -\lambda_y, \end{aligned}\right\} \quad (6\text{-}74)$$

$$\frac{\partial H}{\partial \theta} = 0 = -\frac{\lambda_u F}{m}\sin\theta + \frac{\lambda_v F}{m}\cos\theta \quad (6\text{-}75)$$

$$\frac{\partial G}{\partial \theta_0} - \int_0^T \frac{\partial H}{\partial \theta_0}\, dt = 0 = \gamma_3 V\sin\theta_0 - \gamma_4 V\cos\theta_0 \quad (6\text{-}76)$$

$$\left.\begin{aligned} \frac{\partial G}{\partial x(0)} &= \gamma_1 = \lambda_x(0) \\ \frac{\partial G}{\partial y(0)} &= \gamma_2 = \lambda_y(0) \\ \frac{\partial G}{\partial u(0)} &= \gamma_3 = \lambda_u(0) \\ \frac{\partial G}{\partial v(0)} &= \gamma_4 = \lambda_v(0) \end{aligned}\right\} \quad (6\text{-}77)$$

$$\left.\begin{aligned} \frac{\partial G}{\partial x(T)} &= -\lambda_0 = -\lambda_x(T) \\ \frac{\partial G}{\partial y(T)} &= \lambda_5 = \lambda_y(T) \\ \frac{\partial G}{\partial u(T)} &= 0 = \lambda_u(T) \\ \frac{\partial G}{\partial v(T)} &= 0 = \lambda_v(T) \end{aligned}\right\} \quad (6\text{-}78)$$

$$\frac{\partial G}{\partial T} = 0 = H(T-0) \quad (6\text{-}79)$$

and

$$\lambda_u \cos\Theta + \lambda_v \sin\Theta < \lambda_u \cos\theta + \lambda_v \sin\theta \quad (6\text{-}80)$$

for all admissible Θ.

Eqs. 6-74 yield

$$\begin{aligned} \lambda_x &= \xi_1 \\ \lambda_y &= \xi_2 \\ \lambda_u &= \xi_3 - \xi_1 t \\ \lambda_v &= \xi_4 - \xi_2 t \end{aligned}$$

The last two equations of Eq. 6-78 imply $\xi_3 = \xi_1 T$ and $\xi_4 = \xi_2 T$. If we assume the problem is normal, $\lambda_0 = 1$ so $\lambda_x(T) = 1 = \xi_1$, so

$$\left.\begin{aligned} \lambda_x &= 1 \\ \lambda_y &= \xi_2 \\ \lambda_u &= T \\ \lambda_v &= \xi_2 (T - t) \end{aligned}\right\} \quad (6\text{-}81)$$

Substituting from Eqs. 6-77 and 6-81 into Eqs. 6-75 and 6-76

$$T \sin\theta_0 - \xi_2 T \cos\theta_0 = 0$$

and

$$-(T - t)\sin\theta + \xi_2 (T - t)\cos\theta = 0.$$

For all $t \neq$ T, this is

$$-\sin\theta + \xi_2 \cos\theta = 0.$$

These equations imply

$$\left.\begin{aligned} \xi_2 &= \tan\theta_0 \\ \theta(t) &= \theta_0 \end{aligned}\right\} \quad (6\text{-}82)$$

Integrating the last two equations in Eq. 6-70,

$$\left.\begin{aligned} u(t) &= V\cos\theta_0 + \frac{tF}{m}\cos\theta_0 \\ v(t) &= V\sin\theta_0 + \frac{tF}{m}\sin\theta_0 - gt \end{aligned}\right\} \quad (6\text{-}83)$$

Integrating again,

$$\left.\begin{aligned} x(t) &= tV + \frac{t^2 F}{2m}\cos\theta_0 \\ y(t) &= tV\sin\theta_0 - \frac{1}{2}t^2 g \\ &\quad + \frac{t^2 F}{2m}\sin\theta_0 \end{aligned}\right\} \quad (6\text{-}84)$$

By use of these equations, the last equation in Eq. 6-71 yields

$$T\left(V\sin\theta_0 - \frac{1}{2}gT + \frac{TF}{2m}\sin\theta_0\right) = 0.$$

This implies

$$\sin\theta_0 = \frac{\frac{1}{2}gT}{\left(V + \frac{TF}{2m}\right)}. \quad (6\text{-}85)$$

The one condition which has not been used is Eq. 6-79. By substitution of Eqs. 6-81 and 6-83 into Eq. 6-73, Eq. 6-79 becomes

$$\left(V + \frac{TF}{m}\right)\cos\theta_0 + \xi_2\left(V\sin\theta_0 + \frac{TF}{m}\sin\theta_0 - gT\right) = 0.$$

By use of Eq. 6-82 this becomes

$$\left(V + \frac{TF}{m}\right)\cos^2\theta_0 + \left(V + \frac{TF}{m}\right)\sin^2\theta_0 - gT\sin\theta_0 = 0,$$

or

$$V + \frac{TF}{m} = gT\sin\theta_0. \quad (6\text{-}86)$$

Combining Eqs. 6-85 and 6-86,

$$\left(V + \frac{TF}{m}\right)\left(V + \frac{TF}{2m}\right) = \frac{1}{2}g^2 T^2$$

or

$$\left(\frac{F^2}{2m^2} - \frac{g^2}{2}\right) T^2 + \frac{3}{2}\frac{FV}{m} T + V^2 = 0$$

so

$$T = \frac{-\frac{3}{2}\frac{FV}{m} + \left[\frac{9F^2V^2}{4m^2} - 4V^2\left(\frac{F^2}{2m^2} - \frac{g^2}{2}\right)\right]^{1/2}}{2\left(\frac{F^2}{2m^2} - \frac{g^2}{2}\right)} \tag{6-87}$$

Substituting T from Eq. 6-87 into Eq. 6-85 then gives an easy equation for θ_0.

While the results of this problem are not particularly useful, the solution does illustrate the use of the various conditions in Theorem 6-5 in generating a candidate solution of the problem. The reader, however, should not be led to believe that all Bolza problems may be solved in closed form as in this example.

In more general problems the adjoint equations, Eq. 6-61, cannot be solved so easily in closed form. Further, the equation $\partial H/\partial u = 0$ may not yield so simple a condition as Eq. 6-75 for the design variable. It is often of value to keep a procedure in mind for determining the various unknowns as in this problem, even though more realistic problems may require numerical methods at each step in the procedure.

Example 6-7: Minimum Fuel Orbit Transfer

A rocket equipped with a constant thrust engine is to transfer from a circular earth orbit of radius r_0 to one of radius $R > r_0$ using a minimum of fuel. The time allowed for this transfer is T. Further, it is possible to shut the rocket down during one time interval of the transfer if desired. The orbits are illustrated in Fig. 6-9 and the time scale is shown in Fig. 6-10.

The times t_1 and t_2 may actually coincide, depending on the problem parameters. These times play the role of t^f in the Bolza problem. The equations of motion of the spacecraft are taken as

$$\left.\begin{aligned} \dot{r} &= u \\ \dot{u} &= \frac{v^2}{r} - \frac{\mu}{r^2} + \frac{h(t)F\sin\theta}{m} \\ \dot{v} &= -\frac{uv}{r} + \frac{h(t)F\cos\theta}{m} \\ \dot{m} &= h(t)\,q \end{aligned}\right\} \tag{6-88}$$

where

$$h(t) = \begin{cases} 1, 0 < t < t_1 \\ 0, t_1 < t < t_2 \\ 1, t_2 < t < T \end{cases}$$

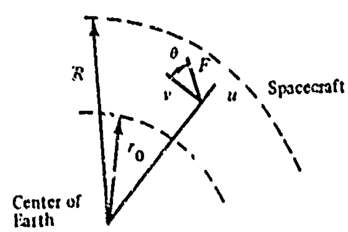


Figure 6-9. Orbit Transfer

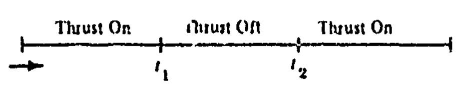


Figure 6-10 Thrust Program

and the boundary conditions are

$$\left.\begin{aligned} &r(0) = r_0,\ u(0) = 0 \\ &v(0) = (\mu/r_0)^{1/2},\ m(0) = m_0, \\ &\qquad r(T) = R,\ u(T) = 0 \\ &v(T) = (\mu/R)^{1/2} \end{aligned}\right\} \tag{6-89}$$

where

r = radius

u = radial velocity

v = tangential velocity

m = mass of spacecraft

μ = gravitation constant

F = thrust

q = mass flow rate during thrust

ϕ = thrust orientation

Since $m_0 - m(t)$ is the amount of fuel consumed up to time t, the object here is to minimize

$$J = m_0 - m(T).$$

For use in Theorem 6-5, define

$$\begin{aligned} G = m_0 &- m(T) + \gamma_1 [r(0) - r_0] + \gamma_2 u(0) \\ &+ \gamma_3 \left[v(0) - (\mu/r_0)^{1/2}\right] \\ &+ \gamma_4 [m(0) - m_0] + \gamma_5 [r(T) - R] \\ &+ \gamma_6 u(T) + \gamma_7 \left[v(T) \quad (\mu/R)^{1/2}\right] \end{aligned}$$

and

$$\begin{aligned} H = \lambda_r u &+ \lambda_u \left[\frac{v^2}{r} - \frac{\mu}{r^2} + \frac{h(t)F\sin\phi}{m}\right] \\ &+ \lambda_v \left[-\frac{uv}{r} + \frac{h(t)F\sin\phi}{m}\right] \\ &+ \lambda_m h(t) q \end{aligned}$$

The necessary conditions of Theorem 6-5 are

$$\left.\begin{aligned} \dot{\lambda}_r &= -\lambda_u \left(-\frac{v^2}{r^2} + \frac{3\mu}{r^2}\right) - \lambda_v \left(\frac{uv}{r^2}\right) \\ \dot{\lambda}_u &= -\lambda_r + \lambda_v \frac{v}{r} \\ \dot{\lambda}_v &= -\lambda_u \left(\frac{2v}{r}\right) + \lambda_v \left(\frac{\text{[illegible]}}{r}\right) \\ \dot{\lambda}_m &= \lambda_u \left[\frac{h(t)F\sin\phi}{m^2}\right] \\ &\quad + \lambda_v \left[\frac{h(t)F\cos\phi}{m^2}\right] \end{aligned}\right\} \tag{6-90}$$

$$\begin{aligned} 0 = \lambda_u &\left[\frac{h(t)F\cos\phi}{m}\right] \\ &- \lambda_v \left[\frac{h(t)F\cos\phi}{m}\right] \end{aligned} \tag{6-91}$$

$$\lambda_m(T) = -1 \tag{6-92}$$

$$\left.\begin{aligned} \lambda(t_1 - 0) &= \lambda(t_1 + 0) \\ \lambda(t_2 - 0) &= \lambda(t_2 + 0) \end{aligned}\right\} \tag{6-93}$$

$$\left.\begin{aligned} H(t_1 - 0) &= H(t_1 + 0) \\ H(t_2 - 0) &= H(t_2 + 0) \end{aligned}\right\} \tag{6-94}$$

the t_i are those shown in Fig. 6-10.

The prospect of solving this set of equations in closed form is dim indeed. A general procedure can be discussed, however, and the actual solution can be obtained using numerical methods discussed in a later paragraph.

From Eq. 6-91, $\phi(t)$ may be determined as

$$\phi(t) = \text{Arctan}\left(\frac{\lambda_u}{\lambda_v}\right) \tag{6-95}$$

for t not in $t_1 \leq t \leq t_2$, where ϕ need not be defined. The result of Eq. 6-95 may be substituted into Eqs. 6-88 and 6-90 so that these equations become a set of eight first-order differential equations for the state and adjoint variables. Eqs. 6-89 and 6-92 form a set of eight boundary conditions for these variables. Eqs. 6-93 show that the adjoint variable is continuous and Eqs. 6-94 determine t_1 and t_2. A numerical procedure may be used to solve this problem. The resulting adjoint variables may then be substituted into Eq. 6-95 to obtain the explicit design (or control) variable. A problem of this kind is discussed in Ref. 18. The method used there to construct a solution is completely different from the one proposed here.

6-4 PROBLEMS OF OPTIMAL DESIGN AND CONTROL

The Bolza problem of par. 6-3 is of almost the generality required for optimal design. The principal shortcoming of that problem is in the lack of generality in the constraints. It has been noted in preceding chapters that meaningful optimal design problems generally involve inequality constraints. It is the purpose of this paragraph to extend the Bolza problem to account for inequality constraints.

The problem treated here is given in Definition 6-3.

Definition 6-3 (Problem of Optimal Design): The optimal design problem is a problem of finding $u(t)$, b, $x(t)$, $t^0 \leq t \leq t^n$, which minimize

$$J = g_0(b, t^j, x^j) + \int_{t^0}^{t^n} f_0\,[t, x(t), u(t), b]\,dt \tag{6-96}$$

subject to the conditions

$$\frac{dx}{dt} = f(t, x, u, b),\ t^0 \leq t \leq t^n,\ t \neq t^j \tag{6-97}$$

$$g_\alpha(b, t^j, x^j) + \int_{t^0}^{t^n} L_\alpha\,[t, x(t), u(t), b]\,dt = 0,\ \alpha = 1, \ldots, r' \tag{6-98}$$

$$g_\alpha(b, t^j, x^j) + \int_{t^0}^{t^n} L_\alpha\,[t, x(t), u(t), b]\,dt \leq 0,\ \alpha = r' + 1, \ldots, r \tag{6-99}$$

$$\left.\begin{array}{l}\phi_\beta(t, x, u, b) = 0,\ \beta = 1, \ldots, q', \\ t^0 \leq t \leq t^n\end{array}\right\} \tag{6-100}$$

and

$$\left.\begin{array}{l}\phi_\beta(t, x, u, b) \leq 0,\ \beta = q' + 1, \ldots, q, \\ t^0 \leq t \leq t^n.\end{array}\right\} \tag{6-101}$$

The variables and functions appearing here are identical to those in Def. 6-2.

The inequalities in this problem are treated here in the manner presented in Ref. 13. The inequality constraints are first reduced to equality constraints, and the results of par. 6-3 are applied. In order to perform this reduction, define the slack variables ν_α, $\alpha = r' + 1, \ldots, r$ and $w_\beta(t)$, $\beta = q' + 1, \ldots, q$ by

$$g_\alpha(b, t^l, x^l) + \int_{t^o}^{t^1} L_\alpha\,[t, x(t), u(t), b]\,dt + \nu_\alpha^2 = 0,\ \alpha = r' + 1, \ldots, r \tag{6-102}$$

and

$$\left.\begin{aligned} &\phi_\beta(t, x, u, b) + w_\beta^2(t) = 0, \\ &\beta = q' + 1, \ldots, q \end{aligned}\right\} \tag{6-103}$$

The constraints, Eqs. 6-102 and 6-103, are equivalent to Eqs. 6-99 and 6-101, respectively, where ν_α and $w_\beta(t)$ are interpreted as design parameters and design variables. With these equality constraints replacing the inequality constraints, the optimal design problem becomes a Bolza problem. The necessary conditions of par. 6-3, therefore, may be applied to this modified problem.

The form of the constraints, Eq 6-101, has a great deal to do with the behavior of the problem. If some function ϕ_β depends only on t, x, and b then the problem is complicated in intervals in which $\phi_\beta = 0$. This kind of constraint will be referred to as a state variable inequality constraint and will be treated separately. If ϕ_β does depend explicitly on u, then the constraint is referred to as a design variable inequality constraint. This problem will now be investigated.

6-4.1 DESIGN VARIABLE INEQUALITY CONSTRAINTS

In order to apply Theorem 6-5 to the problem of Eqs 6-96, 6-97, 6-98, 6-100, 6-102, and 6-103, the independence of conditions expressed by Eqs 6-100 and 6-103 must be verified, i.e., the matrix, Eq 6-55, is required to have rank q. For this purpose, the design vector must be considered as $(u^T, w^T)^T$, where $w^T = (w_{q'+1}, \ldots, w_q)$.

The matrix, Eq. 6-55, becomes

$$\begin{bmatrix} \dfrac{\partial\phi_1}{\partial u_1}, \ldots, \dfrac{\partial\phi_1}{\partial u_m}, 0, \ldots, 0 \\ \vdots \quad\quad \vdots \\ \dfrac{\partial\phi_{q'}}{\partial u_1}, \ldots, \dfrac{\partial\phi_{q'}}{\partial u_m}\quad 0, \ldots, 0 \\ \dfrac{\partial\phi_{q'+1}}{\partial u_1}, \ldots, \dfrac{\partial\phi_{q'+1}}{\partial u_m}, 2w_{q'+1}, \ldots, 0 \\ \vdots \quad\quad \vdots \quad\quad \ddots \\ \dfrac{\partial\phi_q}{\partial u_1}, \ldots, \dfrac{\partial\phi_q}{\partial u_m}, 0, \ldots, 2w_q \end{bmatrix} \tag{6-104}$$

This matrix is required to have rank q. In order for this to be possible the number of columns, $m + q - q'$, must be greater than or equal to the number of rows, q, or $m - q' > 0$. Further, it is obvious that the first q' rows must be linearly independent, or the entire matrix could not possibly have rank q. Next note that if $w_\alpha \neq 0$, then the αth row must be linearly independent of all the other rows since it has the only nonzero element in the $m + \alpha - q'$th column. Therefore, linear independence of the rows from $q' + 1$ to q need only be considered for those α with $w_\alpha = 0$. By Eq. 6-103, this is the same as $\phi_\alpha = 0$.

The conclusion is, then, that the matrix, Eq 6-104, will have rank q if and only if the matrix

$$\left[\frac{\partial\phi_i}{\partial u}\right], \phi_i = 0 \tag{6-105}$$

is of full row rank. This simply says the gradients of all constraint functions (which are equalities) with respect to the design variable must be linearly independent. Assuming this is the case, Theorem 6-5 may be applied.

Define

$$G = \lambda_0 g_0 + \sum_{\alpha=1}^{r} \gamma_\alpha g_\alpha \tag{6-106}$$

$$G' = \sum_{\alpha=r'+1}^{r} \gamma_\alpha \nu_\alpha^2 \tag{6-107}$$

$$H = \lambda^T f - \lambda_0 f_0 - \sum_{\alpha=1}^{r} \gamma_\alpha L_\alpha - \sum_{\beta=1}^{q} \mu_\beta \phi_\beta \tag{6-108}$$

$$H' = -\sum_{\alpha=r'+1}^{r} \gamma_\alpha \nu_\alpha^2 - \sum_{\beta=q'+1}^{q} \mu_\beta w_\beta^2 . \tag{6-109}$$

The quantities $\bar{H} = H + H'$ and $\bar{G} = G + G'$ take the place of H and G in Theorem 6-5. Necessary conditions for the optimal design problem are, therefore,

$$\frac{d\lambda}{dt} = -\frac{\partial \bar{H}^T}{\partial x} \tag{6-110}$$

$$\frac{\partial H}{\partial u} = 0 \tag{6-111}$$

$$\frac{\partial H'}{\partial w} = 0 \tag{6-112}$$

$$\frac{\partial G}{\partial b} + \int_{t^0}^{t^n} \frac{\partial H}{\partial b}\, dt = 0 \tag{6-113}$$

$$\frac{\partial G'}{\partial \nu} + \int_{t^0}^{t^n} \frac{\partial H'}{\partial \nu}\, dt = 0 \tag{6-114}$$

and the conditions, Eqs. 6-110 through 6-113, are unchanged. Further, Theorem 6-6 yields

$$\bar{H}[t,x(t),U,b,\lambda(t),\gamma,0,\nu,w] \leq \bar{H}[t,x(t),u(t),b,\lambda(t),0,\nu,w] \tag{6-115}$$

for all admissible U.

The condition, Eq. 6-112, in scalar form is

$$-2\mu_\beta w_\beta = 0, \beta = q' + 1, ..., q . \tag{6-116}$$

If $w_\beta = 0$, then by Eq. 6-103 $\phi_\beta = 0$. If $w_\beta \neq 0$, $\phi_\beta < 0$ and $\mu_\beta = 0$. Therefore, Eq. 6-116 is equivalent to

$$\mu_\beta(t)\phi_\beta(t,x,u,b) = 0, \ \beta = q' + 1, ..., q . \tag{6-117}$$

Condition, Eq. 6-114, is

$$2\gamma_\alpha \nu_\alpha + \int_{t^0}^{t^n} 2\gamma_\alpha \nu_\alpha\, dt = 0,$$

$$\alpha = r' + 1, ..., r$$

or

$$2\gamma_\alpha \nu_\alpha (1 + t^n - t^0) = 0, \ \alpha = r' + 1, ..., r .$$

Since $1 + t^n - t^0 \neq 0$,

$$\gamma_\alpha \nu_\alpha = 0. \ \alpha = r' + 1, ..., r . \tag{6-118}$$

If $\nu_\alpha = 0$, then by Eq. 6-102, the constraint, Eq. 6-99, is an equality. If $\nu_\alpha \neq 0$, then the constraint, Eq. 6-99, is a strict inequality and $\gamma_\alpha = 0$. Therefore, Eq. 6-118 is equivalent to

$$\gamma_\alpha \left\{ g_\alpha(b,t^j,x^j) + \int_{t^0}^{t^n} L_\alpha[t,x(t),u(t),b]\, dt \right\} = 0,$$

$$\alpha = r' + 1, \ldots, r. \tag{6-119}$$

The conditions, Eqs. 6-116 and 6-118, imply $H' = 0$ and $G' = 0$ so that $\bar{H} = H$ and $\bar{G} = G$. The necessary condition, Eq. 6-115, is, therefore, just

$$H[t,x(t),U,b,\lambda(t),\gamma,0] \leq H[t,x(t),u(t),b,\lambda(t),\gamma,0] \tag{6-120}$$

for all admissible U. It is further shown (Refs. 5,10,12) that $\lambda_0 \geq 0$, $\gamma_\alpha \geq 0$, $\alpha = r' + 1, \ldots, r$, and $\mu_\beta(t) \geq 0$, $\beta = q' + 1, \ldots, q$, $t^0 \leq t \leq t^n$.

The conditions obtained through application of Theorem 6-5 to the optimal design problem may now be stated as Theorem 6-7.

Theorem 6-7: If $[x(t), u(t), b, t^j, x^j]$ is a solution of the optimal design problem of Def. 6-3 and if the matrix, Eq. 6-105, has full row rank, then there exist multipliers $\lambda_0 \geq 0$, $\lambda_i(t)$, $i = 1, \ldots, m$, γ_α, $\alpha = 1, \ldots, r$, $\gamma_\alpha \geq 0$, $\alpha = r' + 1, \ldots, r$, $\mu_\beta(t)$, $\beta = 1, \ldots, q$, $\mu_\beta(t) \geq 0$, $\beta = q' + 1, \ldots, q$, not all zero, and functions G and H of Eqs. 6-106 and 6-108 such that

$$\frac{d\lambda}{dt} = -\frac{\partial H^T}{\partial x}, \text{ for } t \neq t^j \tag{6-121}$$

$$\frac{\partial H}{\partial u} = 0, \text{ for } t \neq t^j \tag{6-122}$$

$$\frac{\partial G}{\partial b} \quad \int_{t^0}^{t^n} \frac{\partial H}{\partial b}\, dt = 0 \tag{6-123}$$

$$\left.\begin{aligned} &\frac{\partial G^T}{\partial x^0} - \lambda(t^0) = 0 \\ &\frac{\partial G^T}{\partial x^n} + \lambda(t^n) = 0 \\ &\frac{\partial G^T}{\partial x^j} + \lambda(t^j - 0) - \lambda(t^j + 0) = 0 \end{aligned}\right\} \tag{6-124}$$

$$\left.\begin{aligned} &\frac{\partial G}{\partial t^0} + H(t^0 + 0) = 0 \\ &\frac{\partial G}{\partial t^n} - H(t^n - 0) = 0 \\ &\frac{\partial G}{\partial t^j} - H(t^j - 0) + H(t^j + 0) = 0 \end{aligned}\right\} \tag{6-125}$$

$$H(t^* - 0) - H(t^* + 0) = 0 \tag{6-126}$$

$$\lambda(t^* - 0) - \lambda(t^* + 0) = 0 \tag{6-127}$$

$$\mu_\beta(t)\phi_\beta(t,x,u,b) = 0, \; \beta = 1, \ldots, q \tag{6-128}$$

$$\gamma_\alpha \left\{ g_\alpha(b,t^j,x^j) + \int_{t^0}^{t^n} L_\alpha[t,x(t),u(t),b]\, dt \right\} = 0$$

$$\alpha = 1, \ldots, r \tag{6-129}$$

$$\frac{dH}{dt} = \frac{\partial H}{\partial t}, \text{ for } t \neq t^j \tag{6-130}$$

and

$$H[t,x(t),U,b,\lambda(t),\gamma,0] \leq H[t,x(t),u(t),b,\lambda(t),\gamma,0] \tag{6-131}$$

for all admissible U.

It should be noted that, just as in Theorem 6-5, the number of conditions here is just equal to the number of unknowns, so one might be led to believe that a solution may be found. Existence of a solution is, however, a very difficult question that is treated in Refs. 10 and 11.

6-4.2 STATE VARIABLE INEQUALITY CONSTRAINTS

In many meaningful design problems constraints may involve restrictions only on the state variable. This is the case when some ϕ_β of Eq. 6-101 depends only on t, x, and b. To study this problem, just one such constraint needs to be considered, i.e.,

$$\phi_\beta(t,x,b) \leq 0,\ t^0 \leq t \leq t^n \,. \qquad (6\text{-}132)$$

Let $t^- \leq t \leq t^+$, $t^- < t^+$, be an interval in which ϕ_β of Eq. 6-132 is an equality. It is clear that $\partial\phi_\beta/\partial u = 0$, so the matrix, Eq. 6-105, has a zero row in this interval and hence cannot be of full row rank. Theorem 6-7 cannot be applied directly, so further analysis is required.

In the interval $t^- \leq t \leq t^+$, $\phi_\beta = 0$ so it is necessary that

$$\frac{d\phi_\beta}{dt} = 0 = \frac{\partial\phi_\beta}{\partial t} + \frac{\partial\phi_\beta}{\partial x}\frac{dx}{dt}.$$

From Eq. 6-97, dx/dt may be replaced by f and this relation becomes

$$0 = \frac{d\phi_\beta}{dt} = \frac{\partial\phi_\beta[t,x(t),b]}{\partial t} + \frac{\partial\phi_\beta[t,x(t),b]}{\partial x} f[t,x(t),u(t),b] \,.$$

If the right side of this equation depends explicitly on $u(t)$, then this equation is of the form required in the problem treated in par. 6-4.1. If not, then differentiating through this equation with respect to t and using the chain rule of differentiation

$$0 = \frac{d^2\phi_\beta}{dt^2} = \frac{\partial^2\phi_\beta}{\partial t^2} + 2\frac{\partial^2\phi_\beta}{\partial t\partial x} f + f^T\frac{\partial^2\phi_\beta}{\partial x^2} f + \frac{\partial\phi_\beta}{\partial x}\frac{df}{dt}, \qquad (6\text{-}133)$$

where all the arguments are omitted. If the right side of Eq. 6-133 depends explicitly on $u(t)$ then this equation is of the form treated in par. 6-4.1.

This process continues until

$$0 = \frac{d^{\nu_\beta}\phi_\beta}{dt^{\nu_\beta}}[t,x(t),b] \qquad (6\text{-}134)$$

involves $u(t)$ explicitly in its right side and $u(t)$ can be determined as a function of $x(t)$ and b, as in par. 6-4.1. The integer $\nu_\beta \geq 1$ is defined to be the first integer for which this is true. The constraint, Eq. 6-132, is then called a ν_βth order state variable inequality constraint.

From the theory of ordinary differential equations (Ref. 14), Eq. 6-134 throughout $t^- < t < t^+$ and

$$\phi_\beta[t^-,x(t^-),b] = 0 \qquad (6\text{-}135)$$

$$\frac{d^i\phi_\beta}{dt^i}[t^-,x(t^-),b] = 0,\ i = 1, \ldots, \nu_{\beta-1} \qquad (6\text{-}136)$$

are equivalent to $\phi_\beta = 0$ throughout $t^- \leq$

$t < t^+$. This, of course, requires that ϕ_β have ν_β piecewise continuous derivatives, and f have $\nu_{\beta-1}$ piecewise continuous derivatives in $t^- < t < t^+$. The point t^- plays the role of a t^j in the problem stated earlier in this paragraph.

It will be assumed that when the right side of Eq. 6-134 is used in place of ϕ_β in computing the matrix, Eq. 6-105, this matrix has full row rank. In this case Theorem 6-7 may be employed. To utilize this theorem, define

$$\phi_\beta,\nu_\beta = \begin{cases} \phi_\beta, & \phi_\beta < 0 \\ \dfrac{d^{\nu_\beta}\phi_\beta}{dt^{\nu_\beta}}, & \phi_\beta = 0 \end{cases} \tag{6-137}$$

where $\nu_\beta = 0$ if ϕ_β involves u explicitly,

$$G = \lambda_0 g_0 + \sum_{\alpha=1}^{r} \lambda_\alpha g_\alpha \tag{6-138}$$

$$\tilde{G} = \sum_\beta \sum_{l=0}^{\nu_\beta - 1} \left\{ \tau_{l,\beta}^- \frac{d^l \phi_\beta}{dt^l} [t^-, x(t^-), b] \right\} \tag{6-139}$$

where this sum on β is extended only over the indices associated with state variable inequality constraints, $\tau_{l,\beta}^-$ are multipliers, and

$$\tilde{H} = \lambda^T f - \lambda_0 f_0 - \sum_{\alpha=1}^{r} \gamma_\alpha L_\alpha - \sum_{\beta=1}^{q} \mu_\beta \phi_\beta,\nu_\beta \tag{6-140}$$

With $\bar{G} = G + \tilde{G}$ and $\tilde{H}$ replacing G and H in Theorem 6-7, a set of necessary conditions for this problem are obtained. They are easily computed and are given here as Theorem 6-8

Theorem 6-8: If $[x(t), u(t), b, t^j, x^j]$ is a solution of the optimal design problem with state variable inequality constraints, then there exist multipliers $\lambda_0 > 0$, $\lambda_i(t)$, $i = 1, ..., n$, γ_α, $\alpha = 1, ..., r$, $\gamma_\alpha > 0$, $\alpha = r' + 1, ..., r$, $\mu_\nu(t)$, $\beta = 1, ..., q$, $\mu_\beta(t) > 0$, $\beta = q' + 1, ..., q$, and $\tau_{l,\beta}^-$, $l = 1, ..., \nu_\beta$ and β associated with a state variable constraint and $\phi_\beta = 0$ in $t^- < t < t^+$,

$$\frac{d\lambda}{dt} = -\frac{\partial \tilde{H}^T}{\partial x}, \text{ for } t \neq t^j, t^-, t^+ \tag{6-141}$$

$$\frac{\partial \tilde{H}}{\partial u} = 0, \text{ for } t \neq t^j, t^-, t^+ \tag{6-142}$$

$$\frac{\partial G}{\partial b} + \frac{\partial \tilde{G}}{\partial b} - \int_{t^0}^{t^n} \frac{\partial \tilde{H}}{\partial b}\, dt = 0 \tag{6-143}$$

$$\left.\begin{aligned} &\frac{\partial G^T}{\partial x^0} - \lambda(t^0) = 0 \\ &\frac{\partial G^T}{\partial x^n} + \lambda(t^n) = 0 \\ &\frac{\partial G^T}{\partial x^j} + \lambda(t^j - 0) - \lambda(t^j + 0) = 0 \\ &\frac{\partial \tilde{G}^T}{\partial x^-} + \lambda(t^- - 0) - \lambda(t^- + 0) = 0 \\ &\lambda(t^+ - 0) - \lambda(t^+ + 0) = 0 \end{aligned}\right\} \tag{6-144}$$

$$\left.\begin{aligned} &\frac{\partial G}{\partial t^0} + \tilde{H}(t^0 + 0) = 0 \\ &\frac{\partial G}{\partial t^n} - \tilde{H}(t^n - 0) = 0 \\ &\frac{\partial G}{\partial t^j} - \tilde{H}(t^j - 0) + \tilde{H}(t^j + 0) = 0 \\ &\frac{\partial \tilde{G}}{\partial t^-} - \tilde{H}(t^- - 0) + \tilde{H}(t^- + 0) = 0 \\ &\tilde{H}(t^+ - 0) + \tilde{H}(t^+ + 0) = 0 \end{aligned}\right\} \tag{6-145}$$

$$\bar{H}(t^* - 0) - \bar{H}(t^* + 0) = 0 \qquad (6\text{-}146)$$

$$\lambda(t^* - 0) - \lambda(t^* + 0) = 0 \qquad (6\text{-}147)$$

$$\mu_\beta(t)\phi_\beta(t,x,u,b) = 0, \quad \beta = 1, \ldots, q \qquad (6\text{-}148)$$

$$\gamma_\alpha \left\{ g_\alpha(b,t^l,x^l) + \int_{t^0}^{t^n} L_\alpha[t,x(t),u(t),b]\, dt \right\} = 0,$$

$$\alpha = 1, \ldots, r \qquad (6\text{-}149)$$

$$\frac{d\bar{H}}{dt} = \frac{\partial \bar{H}}{\partial t}, \text{ for } t \neq t^*, t^l \qquad (6\text{-}150)$$

and

$$\bar{H}[t,x(t),U,b,\lambda(t),\gamma,0] \leq \bar{H}[t,x(t),u(t),b,\lambda(t),\gamma,0] \qquad (6\text{-}151)$$

for all admissible U.

The full set of necessary conditions embodied in this theorem is awesome from a computational point of view. The differential equations for x and λ are subject to multipoint boundary conditions that involve a set of undetermined multipliers. In a gross sense, Eqs. 6-147 may be viewed as determining intermediate points in $t^0 \leq t \leq t^n$ and the associated boundary conditions on $x(t)$ and $\lambda(t)$.

Use of the theorem is further complicated by the fact that the design variable may be determined as the solution of Eq. 6-142 which satisfies Eq. 6-151. This means that u will be determined as a function of x, b and all the multipliers The expression for u will generally take different forms in different subintervals of $t^0 \leq t \leq t^n$ and the spacing of these subintervals is not known before the solution is computed. The generality of the problem makes it difficult to discuss all its intracies without resorting to special cases and examples.

6-4.3 APPLICATION OF THE THEORY OF OPTIMAL DESIGN

In order to develop some familiarity with the methods of the preceding subparagraphs, several examples will be treated here. These problems will be idealizations of real-world problems but will illustrate the basic ideas which carry over into more complicated problems.

Example 6-8: Time-optimal Steering of a Ground Vehicle (Ref. 19)

To illustrate the concepts presented in par. 6-4.2, an optimal vehicle steering problem will be solved This problem is chosen because of its clarity of formulation and solution. A ground vehicle (a tractor in this case) is to be steered so that it begins at a given point and is steered so that it reaches a given straight line path in the shortest possible time. The vehicle and the line it is to reach are shown in Fig. 6-11.

Point A, midway between the rear wheels, is located by the coordinates $x_1(t)$ and $x_2(t)$.

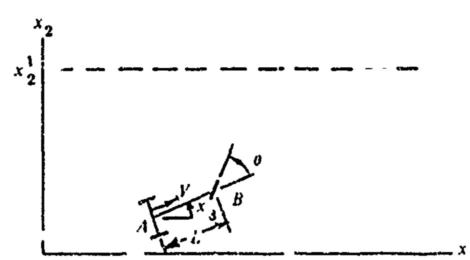


Figure 6-11. Ground Vehicle

Orientation of the vehicle is specified by a third variable $x_3(t)$. Steering of the vehicle is accomplished by choosing the angle $\theta(t)$. From physical grounds it is clear that the state of the vehicle is described by $x(t) = [x_1(t), x_2(t), x_3(t)]^T$ and the vehicle is controlled through choice of $\theta(t)$.

It is assumed that the rear axle of the vehicle moves with a constant velocity V. In this case, motion of the vehicle is governed by the differential equation

$$\left.\begin{aligned} \dot{x}_1 &= V \cos x_3 \\ \dot{x}_2 &= V \sin x_3 \\ \dot{x}_3 &= a \tan \theta \end{aligned}\right\} \qquad (6\text{-}152)$$

where $a = V/L$. At the initial time $t = 0$, $x_1^0(0) = x_1^0$, $x_2(0) = x_2^0$, and $x_3(0) = x_3^0$. The terminal time T is not determined but it is required that $x_2(T) = x_2^1$ and $x_3(T) = 0$ since the vehicle must be tangent to the target line at time T.

The steering angle is limited by

$$-\theta_0 < \theta < \theta_0 \qquad (6\text{-}153)$$

and as an idealization it is assumed that any steering angle in $-\theta_0 < \theta < \theta_0$ may be chosen instantaneously For a reasonable problem it is clear that $\theta_0 < \pi/2$. Further, for definiteness, assume $|x_3^0| < \pi/2$ and $x_2^1 > x_2^0$. All other initial conditions can be obtained from these by reflection in Fig. 6-11.

The problem is now in the form described in par. 6-3. For use in Theorem 6-7,

$$G = \lambda_0 T + \gamma_1 [x_1(0) - x_1^0] + \gamma_2 [x_2(0) - x_2^0] + \gamma_3 [x_3(0) - x_3^0] + \gamma_4 [x_2(T) - x_2^1] + \gamma_5 x_3(T)$$

$$H = \lambda_1 V \cos x_3 + \lambda_2 V \sin x_3 + \lambda_3 a \tan \theta - \mu_1 (\theta - \theta_0) - \mu_2 (\theta_0 - \upsilon).$$

The conditions of Theorem 6-7 are

$$\left.\begin{aligned} \dot{\lambda}_1 &= -\frac{\partial H}{\partial x_1} = 0 \\ \dot{\lambda}_2 &= -\frac{\partial H}{\partial x_2} = 0 \\ \dot{\lambda}_3 &= -\frac{\partial H}{\partial x_3} = \lambda_1 V \sin x_3 \\ &\quad - \lambda_2 V \cos x_3 \end{aligned}\right\} \qquad (6\text{-}154)$$

$$\frac{\partial H}{\partial \theta} = 0 = \lambda_3 a \sec^2 \theta - \mu_1 + \mu_2 \qquad (6\text{-}155)$$

$$\lambda_1(T) = 0 \qquad (6\text{-}156)$$

$$\lambda_0 = \lambda_1(T) V \cos x_3(T) + \lambda_2(T) V \sin x_3(T) + \lambda_3(T) a \tan \theta(T) \qquad (6\text{-}157)$$

$$\left.\begin{aligned} \mu_1(\theta - \theta_0) &= 0 \\ \mu_2(\theta_0 - \theta) &= 0 \end{aligned}\right\} \qquad (6\text{-}158)$$

and

$$\frac{dH}{dt} = \frac{\partial H}{\partial t} = 0. \qquad (6\text{-}159)$$

The first two equations in Eq. 6-154 yield

$$\lambda_1(t) = \xi_1$$

$$\lambda_2(t) = \xi_2$$

and Eq. 6-156 implies $\xi_1 = 0$. The last equation in Eq 6-154 is then

$$\dot{\lambda}_3 = -\xi_2 V \cos x_3 .$$

Using the first equation in Eq. 6-152 to replace $V \cos x_3$, this is

$$\dot{\lambda}_3 = -\xi_2 \dot{x}_1 .$$

Therefore,

$$\lambda_3(t) = -\xi_2 x_1(t) + \xi_3 . \qquad (6\text{-}160)$$

The behavior of $\theta(t)$ may be isolated to two different cases. The first is $|\theta(t)| = \theta_0$. The second is $|\theta(t)| < \theta_0$, in which case Eq. 6-158 implies $\mu_1(t) = \mu_2(t) = 0$. Eq. 6-155 then shows that $\lambda_3(t) = 0$. By Eq. 6-160 then $x_1(t)$ is either a constant or $\xi_2 = \xi_3 = 0$. This and Eq. 6-157 then implies $\lambda_0 = 0$ so all λ_i are zero. This is forbidden by Theorem 6-7, so $x_1(t)$ is a constant when $|\theta(t)| < \theta_0$. But if $x_1(t)$ is constant $\dot{x}_1(t) = 0$ and the first equation in Eq. 6-152 implies $x_3(t) = 0$. The last equation in Eq. 6-152 implies $\theta(t) = 0$.

It is clear then that if $|\theta(t)| < \theta_0$ for some interval of time, the path of the vehicle must be a straight line parallel to the x_2-axis in Fig. 6-11.

Since the last two terms in H are zero, the only explicit dependence of H on θ is through the term $\lambda_3(t) a \tan \theta(t)$. The inequality, Eq. 6-131, states that $\theta(t)$ must maximize H. It is clear then that if $\lambda_3(t) \neq 0$, then

$$\theta(t) = \theta_0 \operatorname{sgn} \{\lambda_3(t)\}, \qquad (6\text{-}161)$$

where

$$\operatorname{sgn} q = \frac{|q|}{q}$$

Further, it is clear that $\theta(t) = 0$ is possible only when $\lambda_3(t) = 0$.

Since $x_3^0 < \pi/2$, for t small, either $\theta(t) = \theta_0$ or $\theta(t) = -\theta_0$. From Fig. 6-11, it is reasonably clear that $\theta(t) = \theta_0$ and Eq. 6-152 can be integrated to obtain

$$\left.\begin{aligned} x_1(t) &= x_1^0 + R\,[\sin(x_3^0 + bt) \\ &\qquad - \sin x_3^0] \\ x_2(t) &= x_2^0 - R\,[\cos(x_3^0 + bt) \\ &\qquad - \cos x_3^0] \\ x_3(t) &= x_3^0 + at \tan\theta_0 , \end{aligned}\right\} \qquad (6\text{-}162)$$

where

$$b = a \tan \theta_0$$

$$R = V/b$$

This path is just a circular arc with center at $(x_1^0 - R \sin x_3^0,\ x_2^0 + R \cos x_3^0)$ and counterclockwise motion.

Similarly, if $\theta(t)$ should become $-\theta_0$ at some time t^* where $x_1(t^*) = x_1^*$, $x_2(t^*) = x_2^*$, and $x_3(t^*) = x_3^*$ then the path is described by

$$\left.\begin{aligned} x_1(t) &= x_1^* - R\,[\sin(x_3^* - bt) \\ &\qquad - \sin x_3^*] \\ x_2(t) &= x_2^* + R\,[\cos(x_3^* - bt) \\ &\qquad - \cos x_3^*] \\ x_3(t) &= x_3^* - at \tan\theta_0 . \end{aligned}\right\} \qquad (6\text{-}163)$$

This path is a circular arc with center at $(x_1^* + R \sin x_3^*,\ x_2^* - R \cos x_3^*)$ and clockwise motion. Since this circular arc must be tangent to the line $x_2 = x_2^1$, the x_2-coordinate of the center must be $x_2^1 - R = x_2^* - R \cos x_3^*$

Note that by Eq. 6-152, $x_3(t)$ must be continuous, so $\dot{x}_1$ and $\dot{x}_2$ are continuous. Therefore, the tangent to the path of point A in the (x_1, x_2)-plane is

$$\frac{dx_2}{dx_1} = \frac{\dot{x}_2}{\dot{x}_1} = \tan x_3$$

and this slope is continuous. This means that segments of the optimal path where $\theta = -\theta_0$, 0, or θ_0 must be tangent where they intersect. With this information, the solution of the problem may be constructed geometrically.

In Fig. 6-12 the initial arc, which is described by Eq. 6-162, is shown leaving (x_1^0, x_2^0). A whole family of second arcs is shown corresponding to different values of x_2^1.

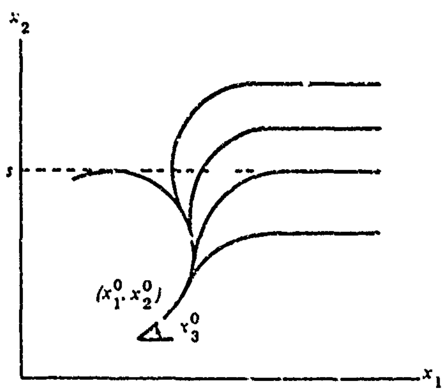


Figure 6-12. Extremal Arcs

From the construction of Fig. 6-12 it is clear that the point of tangency of the two circles (x_1^*, x_2^*) is at the middle of the line joining their centers, i.e.,

$$x_1^* = \frac{1}{2}(x_1^0 - R\sin x_3^0 + x_1^* + R\sin x_3^*)$$

$$x_2^* = \frac{1}{2}(x_2^0 + R\cos x_3^0 + x_2^* - R\cos x_3^*).$$

Further, the relation noted just below Eq. 6-163 is

$$x_2^1 - R = x_2^* - R\cos x_3^*.$$

These equations yield

$$x_1^* = x_1^0 - R\sin x_3^0 + \left[R^2 - \frac{1}{4}(x_2^1 - R - x_2^0 - R\cos x_3^0)^2\right]^{1/2}$$

$$x_2^* = \frac{1}{2}(x_2^0 + x_2^1 + R\cos x_3^0 - R).$$

It may be noted by examining the family of paths in Fig. 6-12 that if $x_2^1 > s = R + x_2^0 + R\cos x_3^0$, then the first arc has been followed beyond a time $\bar{t}$ where $x_3(\bar{t}) = \pi/2$. At the point $x_1(\bar{t}) = x_1^0 - R\sin x_3^0 + R$, $x_2(\bar{t}) = x_2^0 + R\cos x_3^0$ it would have been possible to construct a vertical portion of the optimal path. This construction is shown in Fig. 6-13.

The extremal paths constructed for $x_2^1 > s$ satisfy all the conditions of the theorem so

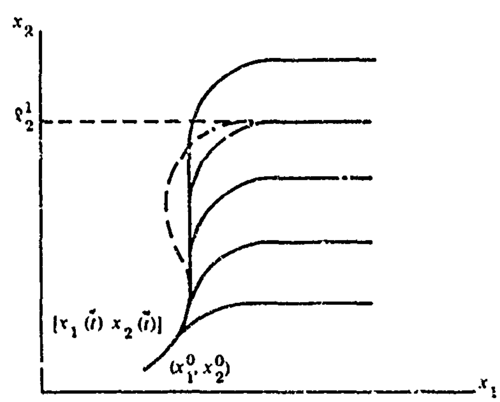


Figure 6-13. Extremal Arcs With Straight Section

that they may be optimum. It is clear that for $x_2^1 < s$ there is only one possible solution of the problem. For $x_2^1 > s$ this is not the case as shown for $x_2^1 = \hat{x}_2^1$. Both the extremals leading to the path $x_2 = \hat{x}_2^1$ satisfy the necessary conditions of the theorem. It is, geometrically, relatively clear that these are the only two possibilities so the one with the shortest time required to get to $x_2 = \hat{x}_2^1$ is to be chosen. The test, Eq. 6-151, may eliminate one candidate. It seems clear that when the extremal with straight line exists, it is best.

It should be noted that if $x_2^1 > s + 2R$, it is impossible to intersect $x_2 = x_2^1$ with only two circular arcs so the extremal with a straight section is required.

This problem illustrates many of the basic ideas and complexities involved in optimal design and optimal control theory. Some of the features are worth noting because they will arise later:

1. Pieced extremals. The conditions of Theorem 6-7 give a set of curves or solutions that must be pieced together to form the optimal path in state space. In the vehicle steering problem, these curves or arcs are put together geometrically. In more complex problems, this will have to be done analytically using the conditions of Theorem 6-7.

2. Multiple solutions. As seen in the foregoing problem, more than one candidate solution may be constructed. Condition, Eq. 6-131, must then be used to choose the best of these candidates.

3. Singular arcs. It occasionally happens, as in the vehicle steering problem, that there will exist a set of values of the state variables and multipliers such that the function H does not depend explicitly on the design variable. In this case, Eq. 6-122 provides no information. It is then required that the inequality, Eq. 6-131, must be used to determine the design variable. For a complete treatment of this subject, see Ref. 20.

The problem treated in this paragraph is not as complicated as most optimal design problems occuring in the real-world. It does, however, illustrate some of the features and difficulty encountered in most realistic optimal design problems. This problem should convince the reader that the solution of optimal design problems is not simply a matter of plugging numbers into formulas. Even though analytical methods will be stressed in subsequent work, the effective solution of this class of problems requires a sound understanding of the theory of optimal design.

Example 6-9: A Constrained Brachistochrone Problem.

The problem considered here is similar to Example 6-4 but with a constraint added. It is required to find the path through (0 0) which lies above the line $x_2 = h + x_1 \tan \alpha$ in the (x_1,x_2)-plane and that carries a particle, without friction, to the vertical line $x_1 = x_1^1$ in the shortest possible time. The problem is shown in Fig. 6-14.

This problem will be treated as an optimal design problem. On the assumption that there are no discontunities in the velocity vector, conservation of energy yields

$$\frac{1}{2}mv^2 = mgx_2$$

or

$$v = (2gx_2)^{1/2}.$$

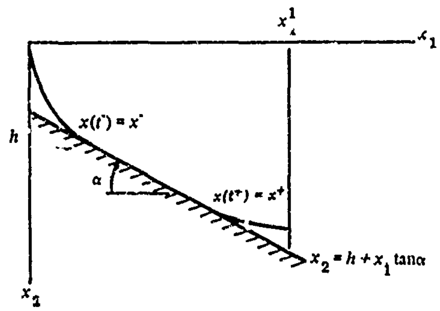


Figure 6-14. Bounded Brachistochrone

The equations of motion of the particle are then

$$\left.\begin{aligned} \dot{x}_1 &= (2gx_2)^{1/2} \cos u \\ \dot{x}_2 &= (2gx_2)^{1/2} \sin u \end{aligned}\right\} \qquad (6\text{-}164)$$

where u is the angle between the x_1-axis and the tangent to the path on which the particle is to travel. This angle u specifies the curve, so it is the design variable. The location of the particle is specified by the point (x_1, x_2) so this is the state variable. The boundary conditions are

$$\left.\begin{aligned} &x_1(0) = x_2(0) = 0 \\ &x_1(T) = x_1^1. \end{aligned}\right\} \qquad (6\text{-}165)$$

The object is then to find $u(t)$, $x_1(t)$, and $x_2(t)$ such that a particle starting at rest at (0,0) reaches $x(T) = x_1^1$ in minimum time T. The path is required to satisfy the constraint

$$\phi = x_2 - x_1 \tan\alpha - h \leq 0. \qquad (6\text{-}166)$$

Since the constraint of Eq. 6-166 does not involve the design variable u explicitly, the problem contains a state variable inequality constraint. Computing $\dot{\phi}$ and substituting from Eq. 6-164 yields

$$\dot{\phi} = (2gx_2)^{1/2} \sin u - (2gx_2)^{1/2} \tan\alpha \cos u$$

$$= 0 \qquad (6\text{-}167)$$

which does contain u explicitly. The constraint, Eq. 6-166, is, therefore, a first-order state variable inequality constraint.

In order to employ Theorem 6-8, define multipliers – γ_i, τ^-, λ_i – such that

$$\left.\begin{aligned} G &= T + \gamma_1 x_1(0) + \gamma_2 x_2(0) \\ &\quad + \gamma_3 [x_1(T) - x_1^1] \\ \bar{G} &= \tau^- (x_2^- - x_1^- \tan\alpha - h) \\ \bar{H} &= \lambda_1 (2gx_2)^{1/2} \cos u \\ &\quad + \lambda_2 (2gx_2)^{1/2} \sin u \\ &\quad \mu (2gx_2)^{1/2} \\ &\quad \times (\cos u - \tan\alpha \sin u). \end{aligned}\right\} \qquad (6\text{-}168)$$

Necessary conditions from Theorem 6-8 are

$$\left.\begin{aligned} \dot{\lambda}_1 &= 0 \\ \dot{\lambda}_2 &= -g(2gx_2)^{1/2} [\lambda_1 \cos u \\ &\quad + \lambda_2 \sin u - \mu(\cos u \\ &\quad \tan\alpha \sin u)] \end{aligned}\right\} \qquad (6\text{-}169)$$

$$(2gx_2)^{1/2} [\quad \lambda_1 \sin u + \lambda_2 \cos u$$

$$+ \mu(\sin u + \tan\alpha \cos u)] = 0 \qquad (6\text{-}170)$$

$$\lambda_2(T) = 0 \qquad (6\text{-}171)$$

$$\left.\begin{aligned} -\tau^- \tan\alpha + \lambda_1 (t^- - 0) \\ -\lambda_1 (t^- + 0) = 0 \\ \tau^- + \lambda_2 (t^- - 0) \\ -\lambda_2 (t^- + 0) = 0 \end{aligned}\right\} \quad (6\text{-}172)$$

$$1 - \tilde{H}(T - 0) = 0 \quad (6\text{-}173)$$

$$-\tilde{H}(t^- - 0) + \tilde{H}(t^- + 0) = 0, \quad (6\text{-}174)$$

and

$$\frac{d\tilde{H}}{dt} = 0. \quad (6\text{-}175)$$

Ideally, the solution for $u(t)$ might proceed by solving Eq. 6-170 for u as a function of λ and μ. This result could then be substituted into Eqs. 6-164, 6-167, and 6-169. The variables λ, x, and μ could then be determined and the results substituted back into the previously derived equation for μ. This would be the desired solution It is clear that these steps would be extremely messy so a huristic argument will be used here to suggest a solution. This solution can then be checked in the conditions Eqs. 6-169 through 6-175.

It might be expected that when $\phi \neq 0$, then the curve is a cycloid as in Example 6-5. Whenever $\phi = 0$ it is clear that $u = \alpha$. This is, in fact, the case and as presented in Ref. 21 the solution is a cycloid for

$$\frac{h}{x_1^1} > \frac{2}{\pi}\left[1 - \left(\frac{\pi}{2} - \alpha\right)\tan\alpha\right],$$

i.e., the optimum path does not touch the constraint surface

For

$$\frac{h}{x_1^1} < \frac{2}{\pi}\left[1 - \left(\frac{\pi}{2} - \alpha\right)\tan\alpha\right]$$

the optimum curve is given by

$$u(t) = \begin{cases} \frac{\pi}{2} - \omega_1 t, & 0 < t < t^- \\ \alpha, & t^- < t < t^+ \\ \omega_2 (T - t), & t^+ < t < T \end{cases}$$

where

$$\omega_1 = \left[\frac{g(\alpha - \pi/2 + \cot\alpha)}{2h\cot\alpha}\right]^{1/2}$$

$$\omega_2 = \left[\frac{g(\alpha + \cot\alpha)}{2(x_1^1 + h\cot\alpha)}\right]^{1/2}$$

$$t^- = \frac{\pi/2 - \alpha}{\omega_1}$$

$$t^+ = T - \alpha/(2\omega_2)$$

and

$$T = \left[\frac{2}{g}(x_1^1 + h\cot\alpha)(\alpha + \cot\alpha)\right]^{1/2} - \left[\frac{2h}{g}\cdot\cot\left(\alpha - \frac{\pi}{2} + \cot\alpha\right)\right]^{1/2}.$$

Fig. 6-15 shows solutions for $\tan\alpha = 1/2$ and several values of h.

The reader may very well get the impression from these examples that analytical solutions of general optimal design problems are extremely difficult to obtain. This is indeed the case. Therefore, either numerical methods must be used to solve the equations given as necessary conditions in the theorems, or some direct computational method must be used to solve the optimal design problem. Some numerical methods of solving the necessary conditions are presented in the next paragraph. Several optimal structural design problems are solved in Chapter 7 to illustrate

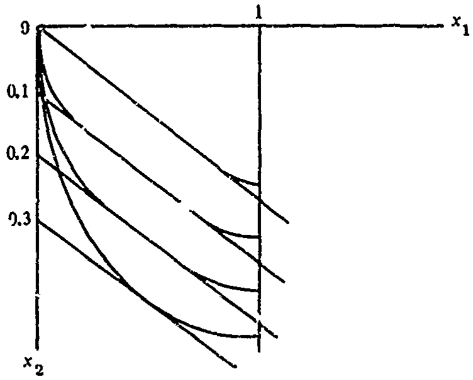


Figure 6-15. Bounded Brachistochrone Solution

these methods. Direct methods of solving optimal design problems are presented in later chapters.

6-5 METHODS OF SATISFYING NECESSARY CONDITIONS

The previous three paragraphs of this chapter have been devoted to obtaining necessary conditions for optimization problems of varying degrees of difficulty. It has been observed that these necessary conditions generally reduce to some sort of boundary-value problem, usually nonlinear. The object of this paragraph is to explore ways in which the boundary-value problem may be solved. This topic has received considerable treatment in the recent literature, so it will be treated only briefly here.

Two different methods will be discussed here and will be applied to optimal structural design problems in the next chapter The first method is based on a reduction of the boundary-value problem to a sequence of initial-value problems whose solutions converge to the solution of the original problem. The second method reduces a nonlinear boundary-value problem to a sequence of linear boundary-value problems whose solutions converge to the solution of the nonlinear problem.

6-5.1 INITIAL VALUE METHODS (OR SHOOTING TECHNIQUES)

In order to develop the main ideas without getting bogged down in notation, consider the problem of finding $y(t) = [y_1(t),..., y_n(t)]^T$, that satisfies

$$\frac{dy}{dt} = f(t,y),\ t^0 < t < t^1 \qquad (6\text{-}176)$$

and

$$\left.\begin{aligned} y_i(t^0) &= y_i^0,\ \text{for some } i \\ y_j(t^1) &= y_j^1,\ \text{for some } j \end{aligned}\right\} \qquad (6\text{-}177)$$

where the total number of conditions in Eq. 6-177 is n. In order to further simplify notation, assume the components of $y(t)$ have been numbered so that the first equation in Eq. 6-177 holds for $i = 1,..., k < n$.

Since initial-value problems are so efficiently integrated forward in time, the missing conditions on y at t^0 may be estimated as

$$y_i(t^0) = \xi_i,\ i = k + 1, ..., n \qquad (6\text{-}178)$$

and Eq. 6-176 integrated from t^0 to t^1 using the full set of initial conditions from the first equation at Eq. 6-177 and Eq 6-178. The value of $y_j(t^1)$ obtained from this integration will probably not satisfy the second equation in Eq 6-177, i.e.,

$$y_j(t^1,\xi) \neq y_j^1,\ j = k + 1, ..., n \qquad (6\text{-}179)$$

where $\xi = (\xi_{k+1}, ..., \xi_n)^T$ and the notation of

Eq. 6-179 is introduced to illustrate the dependence of the final values of y on ξ.

It is clear that a solution of the problem can be obtained if ξ can be found so that Eqs. 6-179 are equalities. To simplify notation, define the column vectors

$\bar{y}(t^1;\xi) = [y_j(t^1;\xi)]$ for those j in Eq. 6-177

and

$\bar{y}^1 = [y_j^1]$ for the same j.

The conditions which are to determine ξ are

$$\bar{y}(t^1;\xi) = y^1. \qquad (6\text{-}180)$$

Any numerical method of solving algebraic equations may be used to solve Eq. 6-189. If a scheme like Newton's Method or a Gradient Method is to be used, it must be possible to compute

$$\frac{\partial \bar{y}}{\partial \xi}(t^1;\xi^0) \qquad (6\text{-}181)$$

where ξ^0 is an estimate of the solution of Eq 6-180. These partial derivatives may be obtained or approximated in a number of ways.

The first method of determining the derivatives in Eq. 6-181 is to observe that $y(t) = y(t;\xi^0)$ and further that the dependence on ξ is very regular (Ref. 4) so that $\partial y(t;\xi^0)/\partial \xi$ exists. Differentiating formally with respect to ξ in Eq. 6-176,

$$\frac{d}{dt}\left(\frac{\partial y}{\partial \xi}\right) = \frac{\partial f}{\partial y}\,\frac{\partial y}{\partial \xi} \qquad (6\text{-}182)$$

and

$$\frac{\partial y}{\partial \xi_i}(0) = [0, ..., \underset{(k)}{0}, .., \underset{(i)}{1}, 0 \,.\, 0],$$

$$i = k+1, .., n \qquad (6\text{-}183)$$

The initial value problems, Eqs. 6-182 and 6-183, for $i = k+1, ..., n$ may be integrated from t^0 to t^1 to obtain the derivatives required in Eq. 6-181.

Once these derivatives have been determined, the new estimate ξ^1 in Newton's Method is given by

$$\xi^1 = \xi^0 - \left[\frac{\partial \bar{y}(t^1;\xi^0)}{\partial \xi}\right]^{-1} [\bar{y}(t^1;\xi^0) - \bar{y}^1]\,. \qquad (6\text{-}184)$$

The process is repeated with ξ^1 playing the role previously occupied by ξ^0.

This method of finding the partial derivatives is direct in nature but requires the solution of Eq. 6-182 $n - k$ times. Further, both the differential equations, Eq. 6-176 and 6-182, must be programmed.

A second method of constructing the partial derivatives of Eq. 6-181 (or approximations of them) is to use a difference quotient, i.e., Eq. 6-176 is solved for ξ and $\xi + \delta$ where $\delta = (0, ..., \underset{i}{\epsilon}, ..., 0)^T$ where i indicates the ith position and ϵ is small. Therefore

$$\frac{\partial y(t^1;\xi)}{\partial \xi_i} \approx \frac{y(t^1;\xi+\delta) - y(t^1;\xi)}{\epsilon} \qquad (6\text{-}185)$$

Once these approximate derivatives are determined, the algorithm, Eq. 6-184, may be used.

This approximate method of constructing the partial derivatives requires that the differential equation, Eq. 6-176, be solved $n - k$ additional times. It, therefore, requires approximately the same amount of computation as the previous scheme, but all the computation is performed with the same set of differential equations. This method is illustrated in the problems of pars. 7-2 and 7-3.

A third scheme which makes use of differentiation formulas for definite integrals is developed in par. 7-4.

6-5.2 A GENERALIZED NEWTON METHOD

A second method which is used to solve the necessary conditions for optimization problems is a Generalized Newton Method of solving boundary-value problems. It has been pointed out in the foregoing that Bolza problems and optimal design problems may be reduced to nonlinear boundary-value problems. The method employed here was developed for just such problems (Refs. 22, 23).

In order to introduce the Generalized Newton Method for boundary-value problems, consider the system of first-order equations

$$\frac{dy}{dt} = g(y,t) \qquad (6\text{-}186)$$

where

$$y(t) = [y_1(t), \ldots, y_n(t)]^T \text{ and}$$

$$g(y,t) = [g_1(y,t), \ldots, g_n(y,t)]^T$$

In addition to satisfying Eq. 6-186, $y(t)$ is required to satisfy

$$\left.\begin{aligned} y_i(t^0) &= y_i^0, \text{ for some } i \\ y_j(t^1) &= y_j^1, \text{ for some } j, \end{aligned}\right\} \qquad (6\text{-}187)$$

where the total number of conditions in Eq. 6-187 is n.

The Generalized Newton Method for solving Eqs. 6-186 and 6-187 is similar in philosophy to the Newton method of solving algebraic equations. An estimate of the solution, $y^{(0)}(t)$, is made and the right side of Eq. 6-186 is expanded about $y^{(0)}(t)$ using Taylor's formula to obtain

$$\frac{dy^{(1)}}{dt} \approx \frac{\partial f}{\partial y}[t,y^{(0)}(t)]y^{(1)} + f[t,y^{(0)}(t)] - \frac{\partial f}{\partial y}[t,y^{(0)}(t)]y^{(0)}(t) \qquad (6\text{-}188)$$

where $y^{(1)}(t)$ is required to satisfy

$$\begin{aligned} y_i^{(1)}(t^0) &= y_i^0, \text{ for those } i \text{ in Eq. 6-187} \\ y_j^{(1)}(t^1) &= y_j^1, \text{ for those } j \text{ in Eq. 6-187} \end{aligned} \qquad (6\text{-}189)$$

The boundary-value problem for $y^{(1)}(t)$ is linear so that if it has a solution, that solution may be obtained by superposition techniques, or any other technique for solving linear boundary-value problems, for that matter (Ref. 24).

The function $y^{(1)}(t)$ is taken as an improved estimate for the solution of Eqs. 6-186 and 6-187. This estimate then replaces $y^{(0)}(t)$ in the preceding analysis. If k is the iteration number for this process, then $y^{(k)}(t)$ is determined by

$$\left.\begin{aligned} \frac{dy^{(k)}}{dt} &= \frac{\partial f}{\partial y}\left[t,y^{(k-1)}(t)\right]y^{(k)} \\ &\quad + f[t,y^{(k-1)}(t)] \\ &\quad - \frac{\partial f}{\partial y}[t,y^{(k-1)}(t)]y^{(k-1)}(t) \end{aligned}\right\} (6\text{-}190)$$

and the boundary conditions

$$y_i^{(k)}(t^0) = y_i^0, \text{ for those } i \text{ in Eq. 6-187}$$

$$y_j^{(k)}(t^1) = y_j^1, \text{ for those } j \text{ in Eq. 6-187}. \tag{6-191}$$

The sequence of approximations to the solution $[y^{(k)}(t)]$ is considered to have converged when the difference between successive iterates is sufficiently small. Theorems given in Ref. 23 show that if the initial estimate of the solution $y^{(0)}(t)$ is sufficiently accurate, then under rather restrictive conditions, the sequence $[y^{(k)}(t)]$ converges to the solution of Eqs. 6-186 and 6-187. Further, the convergence is quadratic in the sense that the error at the $k + 1$st iteration is proportional to the error squared in the kth iteration. This kind of convergence is extremely nice.

Even though it is difficult or impossible to verify the hypotheses of the convergence theorems in Ref. 23, it has been observed in practice (Ref. 23) that good convergence is nevertheless obtained in many real-world problems.

Since the discussion in this paragraph is on ways of solving optimization problems, the Generalized Newton Method will be applied more directly to this class of problems. For the present, consider only the following problem:

$$\text{minimize } J = \int_{t^0}^{t^1} f(t,x,u)dt \tag{6-192}$$

subject to

$$\frac{dx}{dt} = f(t,x,u) \tag{6-193}$$

and

$$\left.\begin{aligned} x_i(t^0) &= x_i^0 \text{ for some } i \\ x_j(t^1) &= x_j^1 \text{ for some } j \end{aligned}\right\} \tag{6-194}$$

where the total number of boundary conditions in Eq. 6-194 may be less than, equal to, or greater than n, $x(t) = [x_1(t) \ldots x_n(t)]^T$, $u(t) = [u_1(t), \ldots, u_m(t)]^T$.

Defining

$$H = \lambda_0 f_0 + \sum_{i=1}^{n} \lambda_i f_i,$$

the necessary conditions of Theorem 6-5 are

$$\frac{d\lambda}{dt} = -\frac{\partial H^T}{\partial x} \tag{6-195}$$

$$\frac{\partial H}{\partial u} = 0 \tag{6-196}$$

and

$$\left.\begin{aligned} \lambda_r(t^0) &= 0, r \neq i \text{ in Eq. 6-194} \\ \lambda_s(t^1) &= 0, s \neq j \text{ in Eq. 6-194}. \end{aligned}\right\} \tag{6-197}$$

The argument used in applying the Generalized Newton Method to the problem of determining $x(t)$, $u(t)$, and $\lambda(t)$ from Eqs. 6-193, 6-194, and 6-195 through 6-197 as follows:

1. Solve Eq. 6-187 for

$$u = u(t,x,\lambda) \tag{6-198}$$

and substitute this expression into Eqs. 6-193 and 6-195.

2. These differential equations then form $2n$ first-order, nonlinear differential equations

in $2n$ variables. Further, there are exactly $2n$ boundary conditions in Eqs. 6-194 and 6-197. This nonlinear boundary-value problem is now solved by the Generalized Newton Method.

3. The solution $x(t)$, $\lambda(t)$ is then substituted into Eq. 6-198 to obtain the optimal design function.

Since the Generalized Newton Method, as presented here, is only capable of solving two-point boundary-value problems, inequality constraints may not be treated explicitly. Rather, the general optimal design problem with inequality constraints must be reduced to a problem with only equality constraints. For example, for problems with constraints of the form

$$\phi_i(t,x,u) < 0, \tag{6-199}$$

where ϕ_i depends explicitly on u a transformation may be performed by introducing an auxiliary design variable (slack variable) $\alpha_i(t)$ through the relation

$$\phi_i(t,x,u) + \alpha_i^2(t) = 0. \tag{6-200}$$

It is clear that with the new variable, Eq. 6-200 is equivalent to Eq. 6-199. The necessary conditions of Theorem 6-5 may now be applied and the Generalized Newton Method utilized just as in the preceding case.

In case the optimal design problems with state variable inequality constraints, a different technique for elimination of inequalities has proved effective. For constraints of the form

$$\psi_i(t,x) < 0 \tag{6-201}$$

an auxiliary parameter ϵ_i is introduced through

$$\int_{t^0}^{t^1} \psi_i^2(t)\, H[\psi_i(t)]\, dt = \epsilon_i \tag{6-202}$$

where

$$H(s) = \begin{cases} 0, s < 0 \\ 1, s > 0. \end{cases}$$

In a sense, ϵ_i is a measure of violation of Eq. 6-201. The procedure in solving an optimal design problem with a constraint of this kind is to solve a sequence of problems with Eq. 6-202 replacing Eq. 6-201, and $\epsilon_i^{(k)}$ approaching zero as k becomes infinite; i.e., a modified design problem is solved imposing Eq. 6-202 in place of Eq. 6-201 with $\epsilon_i^{(0)} > 0$ chosen. This solution is carried out through use of the Generalized Newton Method described. The problem is then solved again with $0 < \epsilon^{(1)} < \epsilon^{(0)}$ beginning the iteration with the solution of the preceding problem. The process is repeated with $0 < \epsilon^{(k)} < \epsilon^{(k-1)}$ until changes in successive solutions are sufficiently small.

The Generalized Newton Method presented here has been discussed by many authors and generally has received favorable comments. For a more detailed discussion and examples, see Refs. 23, 25 through 28. An outstanding treatment of the Generalized Newton Method also appears in book form (Ref. 29). A very rigorous treatment of existence and convergence properties of the method is given which applies to the control problems discussed. The reader should note that some writers follow Bellman in calling the method described here, "Quasilinearization".

REFERENCES

1. D.G. Luenberger, *Optimization by Vector Space Methods*, John Wiley, & Sons, New York, 1969.

2. I.M. Gelfand and S.V. Fomin, *Calculus of Variations*, Prentice-Hall, Englewood Cliffs, New Jersey, 1963.

3. N.I. Akhiezer, *The Calculus of Variations*, Blaisdell, New York, 1962.

4. L.E. Elsgolc, *Calculus of Variations*, Pergamon, London, 1962.

5. M.R. Hesteness, *Calculus of Variations and Optimal Control Theory*, John Wiley & Sons, New York, 1966.

6. L.A. Liusternik and V.J. Sobolev, *Elements of Functional Analysis*, Ungar, New York, 1961.

7. E.K. Blum, "The Calculus of Variations, Functional Analysis, and Optimal Control Problems", in *Topics in Optimization*, G. Leitmann Ed., Academic Press, New York, 1967.

8. M.R. Hesteness, "On Variational Theory and Optimal Control Theory", *J. SIAM*, Series A: *Control*, Vol. 3, No. 1, pp. 23-48, 1965.

9. G.A. Bliss, *Lectures on the Calculus of Variations*, University of Chicago Press, Chicago, 1961.

10. E.B Lee and L. Markus, *Foundations of Optimal Control Theory*, John Wiley & Sons, New York, 1967.

11. W.W. Schmaedke, "The Existence Theory of Optimal Control Systems", in *Advances in Control Systems*, Vol. 3, C.T. Leondes Ed., Academic Press, New York, 1966.

12. L.S. Pontryagin, V.G. Boltyanskii, R.V. Gamkrelidze, and E.F. Mishchenko, *The Mathematical Theory of Optimal Processes*, John Wiley & Sons, New York, 1962.

13. L.D. Berkovitz, "Variational Methods in Problems of Control and Programming", *J. Math. Anal. Appl.*, Vol. 3, pp. 145-169, 1961.

14. E.A. Coddington and N. Levinson, *Theory of Ordinary Differential Equations*, McGraw-Hill, New York, 1955.

15. G. Leitmann, Ed., *Optimization Techniques*, Academic Press, New York, 1962.

16. W.F. Denham, *Steepest-Ascent Solution of Optimal Programming Problems*, Raytheon Report No. BR-2393, Raytheon Company, Bedford, Mass., 1963.

17. A.V. Balakrishnan and L.W. Neustadt, Ed., *Computational Methods in Optimization Problems*, Academic Press, New York, 1964.

18. D.S. Hauge, "Solution of Multiple-arc Problems by the Steepest-Descent Method", in *Recent Advances in Optimization Techniques*, A. Lavi and T.P. Vogl, Eds., John Wiley & Sons, New York, 1966.

19. H.M. Hung, *Time-Optimal Steering of a Ground Vehicle*, presented at the 1968 Spring Meeting of SIAM, Toronto, Canada.

20. H.J. Kelly, R.E. Kopp, and H.G. Moyer, "Singular Extremals", in *Topics in Optimization*, G. Leitmann, Ed., Academic Press, New York, 1967.

21. A.E. Bryson Jr., W.F. Denham, and S.E. Dreyfus, "Optimal Programming Problems with Inequality Constraints I: Necessary Conditions for Extremal Solutions", *AIAA J.*, Vol. 11, November 1963, pp. 2544-2550.

22. L.V. Kantorovich and G.P. Akilov, *Functional Analysis in Normed Spaces*, Pergamon Press, London, 1965.

23. P. Kenneth and R. McGill, "Two-Point Boundary-Value-Problems Techniques", in *Advances in Control Systems*, Vol. 3, C.T. Leondes, Ed., Academic Press, New York, 1966.

24. H.B. Keller, *Numerical Methods for Two-Point Boundary-Value Problems*, Blaisdell, Waltham, Mass., 1968.

25. G. Paine, *The Application of the Method of Quasilinearization to the Computation of Optimal Control*, Report No. 67-49, Department of Engineering, UCLA, August 1967.

26. R.E. Kopp and H.G. Moyer, "Trajectory Optimization Techniques", in *Advances in Control Systems*, Vol. 4, C.T. Leondes, Ed., Academic Press, New York, 1966.

27. A.P. Sage, *Optimum Systems Control*, Prentice-Hall, Englewood Cliffs, N.J., 1968.

28. B.D. Tapley and J.M. Lewallen, "Comparison of Several Numerical Optimization Methods", *J. Opt. Theory and Appl.*, Vol. 1, No. 1, 1967, pp. 1-32.

29. P.L. Falb, *Some Successive Approximation Methods in Control and Oscillation Theory*, McGraw-Hill, New York, 1971.

CHAPTER 7

OPTIMAL STRUCTURAL DESIGN BY THE INDIRECT METHOD

7-1 INTRODUCTION

7-1.1 THE CLASS OF PROBLEMS CONSIDERED

Since the beginning of engineering disciplines, the engineer has attempted to develop structures and machines that perform some specified task. In the case of structures, a frame or truss is required to support a given system of loads Likewise, machines and machine elements are required to support loads while they perform some function.

The objective of the examples treated here is to illustrate organized methods that the engineer may use to obtain a load-carrying system which is best in some sense that is associated with the particular application. In design of commercial goods, the dollar cost of an element is probably the index that is to be minimized (Ref. 1). In military and aerospace applications, while dollar cost is important, frequently weight cost is even more essential. In the example problems presented here, the criterion of minimum weight will be chosen.

Until very recently, most design procedures depended on the engineer's intuition and experience in proportioning a load-carrying system. An analysis of the proposed configuration was then made to determine whether the system met all requirements placed on it. If not, or if the preliminary design was obviously excessively strong, the procedure was repeated until a satisfactory solution was obtained.

As systems become more complex and more emphasis is placed on minimum cost, the designer is unable to make all the trade-off analyses mentally. A method of design synthesis, therefore, is necessary which is able to include all requirements on the system and the requirement of minimum cost in a unified design procedure. One such method for optimal structural design is illustrated in this chapter.

7-1.2 HISTORICAL DEVELOPMENT

Very early in the development of mechanics of materials, methods of determining stress and displacement for given bodies under the action of given forces were emphasized. As these methods became better developed, the question arose as to how a structure might be proportioned to satisfy certain requirements and be best in some sense. Problems of this kind were considered by Lagrange (Ref 2) in 1771 and by Clausen (Ref. 3) in 1851.

Until very recent years, methods of the calculus of variations were not sufficient for treating realistic design problems. Probably for this reason, design problems were stated in terms of a few parameters that specified the structure. For example, uniform beams of undetermined depth are placed in a given configuration. The depths are then determined so that the structure supports the given loads and is as light as possible. For a detailed bibliography of this development through 1963, see Ref. 4. For a more current bibliography, see Ref. 5.

Another important method of design developed principally by Prager and Drucker (Refs. 6,7,8) is limit analysis. In this method of design, the structure is allowed to reach a state of collapse due to plastic action of the material. The resulting design is, therefore, safe for application of the given loads even though permanent deformation of the structure results. If the loads must be applied many times in the life of the structure, it will generally be required that all material in the structure must remain in the elastic range at all times. For this reason, methods had to be developed for elastic design.

In 1960, Joseph B. Keller published an article on column design (Ref. 9) which renewed interest in elastic, minimum weight design. Several papers have subsequently been published by Keller and his associates in which a class of eigenvalue problems is treated (Refs. 10,11,12). The methods employed in these papers are elegant but are not easily adapted to realistic engineering problems.

A new method of optimal design has been developed by J.E. Taylor and W. Prager since 1967 (Refs. 13,14,15). This method is based on an energy representation of the structural element under consideration. A particularly nice feature of the method is the ability to obtain sufficient conditions for certain classes of design problems. However, no unified method of constructing solutions has been presented.

7-1.3 METHODS EMPLOYED

The theorems of Chapter 6 will be employed here for the solution of optimal design problems. Use of the results of Chapter 6 to construct solutions of optimal design problems is called an *indirect method* of solution. This is so, because one first obtains a set of conditions that the solution of the optimal design problems must satisfy. Once this task is complete, the design problem is reduced to the determination of solutions of the necessary conditions that are candidate solutions of the optimal design problem. The term "indirect" seems to describe this process quite well.

As discussed in par. 6-5, any method of solving the nonlinear boundary-value problem contained within the necessary conditions is admissible. In this chapter, two problems will be solved by shooting techniques. The problems of par. 7-2 are treated by the shooting technique of par. 6-5. The problems of par. 7-3, however, are treated by a modified shooting technique.

7-2 A MINIMUM WEIGHT COLUMN

A lightweight column of length T is to be designed to support a given load P. The material is specified and has yield strength σ_{max}. The particular support considered is shown in Fig. 7-1. In problems considered

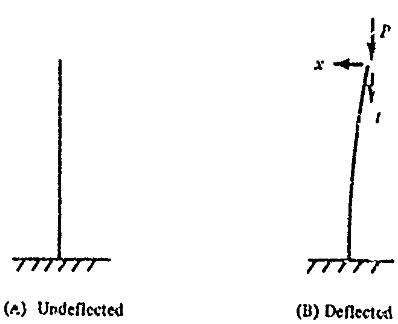


Figure 7-1. Column Under Consideration

here, the cross section is assumed to depend on only one design variable, $u(t)$, $0 < t < T$. The problem is to determine $u(t)$ that minimizes the weight or, equivalently, the volume

$$J = \int_0^T A[u(t)]\,dt \tag{7-1}$$

and satisfies the conditions,

$$\left.\begin{aligned} &EI(u)\frac{d^2x}{dt^2} + Px = 0, \\ &x(0) = 0, \frac{dx}{dt}(T) = 0 \end{aligned}\right\} \tag{7-2}$$

and

$$\phi = \frac{P}{A(u)} - \sigma_{max} \leq 0 \tag{7-3}$$

where

$x(t)$ = lateral deflection of the column

t = distance measured along the column

$A(u)$ = area of the cross section

$I(u)$ = smallest moment of inertia of the area of the cross section about a centroidal axis. All cross sections are assumed to have two orthogonal axes of symmetry with P acting through their intersection.

By defining $x_1 = x$ and $x_2 = dx_1/dt$, Eq. 7-2 reduces to the system

$$\left.\begin{aligned} &\frac{dx_1}{dt} = x_2 \equiv f_1 \\ &\frac{dx_2}{dt} = -\frac{Px_1}{EI(u)} \equiv f_2 \\ &x_1(0) = 0, x_2(T) = 0 . \end{aligned}\right\} \tag{7-4}$$

The problem is thus reduced to the form of the optimal control problem considered in par. 6-4.

For use in Theorem 6-7, construct

$$H = -\lambda_0 A(u) + \lambda_1 x_2 - \lambda_2 \left[\frac{Px_1}{EI(u)}\right] - \mu\left[\frac{P}{A(u)} - \sigma_{max}\right]$$

$$G = \lambda_1 x_1(0) + \lambda_2 x_2(T).$$

Conditions, Eqs. 6-121 and 6-124, yield

$$\left.\begin{aligned} &\frac{d\lambda_1}{dt} = -\frac{\partial H}{\partial x_1} = \frac{P\lambda_2}{EI(u)} \\ &\frac{d\lambda_2}{dt} = -\frac{\partial H}{\partial x_2} = -\lambda_1 \\ &\lambda_2(0) = 0, \lambda_1(T) = 0 . \end{aligned}\right\} \tag{7-5}$$

The system, Eq. 7-5, reduces to

$$\left.\begin{aligned} &\frac{d^2\lambda_2}{dt^2} = -\frac{P\lambda_2}{EI(u)} \\ &\lambda_2(0) = 0, \frac{d\lambda_2}{dt}(T) = 0 . \end{aligned}\right\} \tag{7-6}$$

Eq. 7-6 for $\lambda_2(t)$ is identical to Eq. 6-14 for $x(t)$. Both problems are homogeneous, however, so $\lambda_2(t)$ and $x(t)$ may differ by an arbitrary constant multiplier, say λ_0, i.e., put $\lambda_2(t) = \lambda_0 x(t)$ This problem is normal (Ref. 17), so $\lambda_0 \neq 0$ may be chosen as one.

Condition, Eq. 6-128, of Theorem 6-7 is, in this case, $\mu(P/A(u) - \sigma_{max}) = 0$. Two possibilities now exist: either $\mu = 0$, or $P/A(u) - \sigma_{max} = 0$. In the second case, u is just the algebraic solution of

$$P/A(u) = \sigma_{max}. \tag{7-7}$$

In the remaining case, $\mu = 0$ and condition, Eq. 6-122, of Theorem 6-7 is

$$\frac{\partial H}{\partial u} = -\frac{\partial A}{\partial u} - \frac{Px^2}{E}\left\{\frac{\partial}{\partial u}[1/I(u)]\right\} = 0. \tag{7-8}$$

The design variable $u(t)$ is thus determined in subintervals of $[0,T]$ by either Eq. 7-7 or 7-8. So that the results of the present method may be compared with those obtained by Keller (Ref. 9), choose $A = u$ and $I = \alpha u^2$. This corresponds to having the geometric shape of the cross section fixed and allowing all dimensions to vary as $u^{1/2}$.

With this form of $A(u)$ and $I(u)$, Eqs. 7-7 and 7-8 become

$$P/u = \sigma_{max} \tag{7-9}$$

and

$$1 - \frac{2Px^2}{E\alpha u^3} = 0. \tag{7-10}$$

Condition, Eq. 6-131, of Theorem 6-7 requires that the expression for u be chosen which satisfies the constraint, Eq. 7-3, and makes H as large as possible. This criterion yields the choice between

$$u(t) = \left[(P/\sigma_{max}) \text{ or } \left(\frac{2P}{E\alpha}\right)^{1/3} x^{2/3}\right] \tag{7-11}$$

When Eq. 7-11 is substituted into Eq. 7-2,

$$\frac{d^2x}{dt^2} = -\frac{Px}{E\alpha}\left[(\sigma_{max}/P)^2 \text{ or } \left(\frac{E\alpha}{2P}\right)^{2/3} x^{-4/3}\right] \tag{7-12}$$

$$x(0) = 0, \frac{dx}{dt}(T) = 0$$

where the choice on the right side of Eq. 7-12 must correspond to the selection in Eq. 7-11.

This boundary-value problem is solved by an iterative method based on Newton's Algorithm. The missing initial condition is taken as $dx/dt(0) = C$. Integration of the resulting initial value problem from 0 to T yields an error $dx/dt(T;C)$ in the final value.

This notation is chosen to emphasize the dependence of x on the estimate C of the missing initial condition. The objective is to find C so that $dx/dt(T;C) = 0$. Once C is found, the initial value problem for $x(t)$ may be solved and $u(t)$ determined from Eq. 7-11.

In order to employ Newton's Algorithm, $\partial/\partial C\,[dx/dt(T;C)]$ is needed. It is obtained by formally differentiating Eq. 7-12 with respect to C to obtain

$$\left.\begin{aligned} \frac{d^2\zeta}{dt^2} = -\frac{\sigma^2_{max}}{E\alpha P} &\text{ or } \\ \frac{1}{3}\left(\frac{P}{4E\alpha}\right)^{1/3} & x^{-4/3}\,\zeta \\ \zeta(0) = 0, \frac{d\zeta}{dt}&(0) = 1 \end{aligned}\right\} \tag{7-13}$$

where the choice on the right side of Eq. 7-13 must correspond to the selection in Eq. 7-11. The order of taking derivatives has been changed and the notation $\zeta = \partial x/\partial C$ introduced in obtaining Eq. 7-13.

The iterative method for determining C is then:

Step 1. Make estimate $C = C_0$

Step 2. Integrate the differential equation in Eq. 7-12 with

$$\frac{dx}{dt}(0) = C_0,\ x(0) = 0, \text{ and Eq. 7-13.}$$

Step 3. Make the adjustment in C

$$C_1 = C_0 - \left[\frac{dx}{dt}(T) \Big/ \frac{d\xi}{dt}(T)\right]$$

Step 4. Return to Step 2 with new estimate $C = C_1$, and repeat.

The equations derived here must be changed only slightly to solve column problems with other end conditions and other forms of cross section.

For a numerical example of this problem, let the cross section be circular with variable radius. In this case $\alpha = 1/(4\pi)$. For the example, let $\sigma_{max} = 20{,}000$ psi, $E = 3\times10^7$ psi, and $T = 10$ in.

A FORTRAN program was written to perform the iterative procedure. The program was run on an IBM 360-65 Computer and required approximately 0.1 sec per iteration and only four to six iterations to converge.

The results for this design problem are given in Table 7-1 and Fig. 7-2. For loads above 6794 lb, the cross-sectional area is determined by $A = P/\sigma_{max}$ and the resulting column is stable. A meaningful optimal design problem then exists only for $P < 6794$.

7-3 A MINIMUM WEIGHT STRUCTURE WITH ANGULAR DEFLECTION REQUIREMENTS

7-3.1 STATEMENT OF THE PROBLEM

The problem considered in this paragraph is the design of a portable communication tower of height L which will support a line-of-site transmission unit, a laser transmitter, for example. In order for the transmission beam

TABLE 7-1

RESULTS FOR COLUMN PROBLEM

P, lb	Volume, in.^{3}	Volume of Uniform Column[1], in.^{3}	Saving, %
50	0.260	0.291	10.6
100	0.361	0.412	12.4
200	0.507	0.595	14.7
500	0.806	0.923	12.7
1000	1.140	1.300	12.3
1500	1.408	1.600	11.9
2000	1.640	1.840	10.8
3000	2.048	2.260	9.3
4000	2.412	2.600	7.3
5000	2.765	2.910	5.2
6794	3.397	3.397	0.0

[1] Minimum Weight Uniform Column

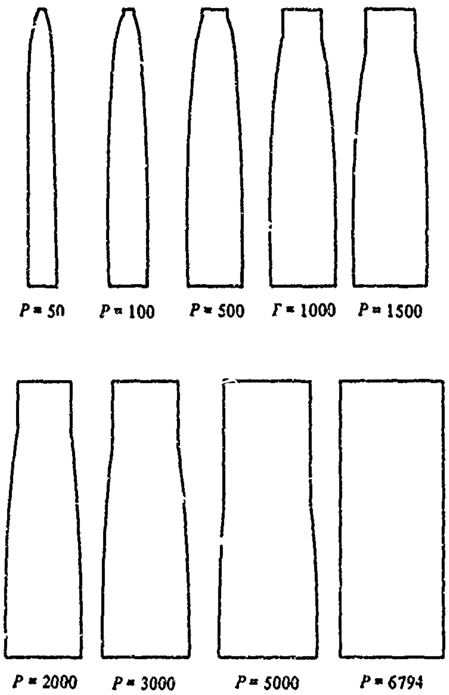


Figure 7-2. Profiles of Optimal Columns

to hit the receiving unit, the top of the tower on which the transmitter unit is mounted must undergo only a certain allowable rotation θ when the tower is exposed to a given extreme uniform wind load Q pounds per unit length of tower.

It is required that the tower be as lightweight as possible so that it may be transported and erected without the aid of heavy machinery. For this reason, the design criterion is minimum weight. However, one additional requirement must be placed on the tower. In transportation and erection it must be strong enough so that it is not damaged by rough treatment. Therefore, it is required that the moment of inertia of the cross-sectional area of the tower be greater than a predetermined limit I_0 everywhere.

The general configuration of the tower is shown in Fig. 7-3. Three vertical members with cross-sectional area $A(t)$ are arranged on the vertices of an equilateral triangle of altitude $h(t)$. Here t is a coordinate measured along the length of the tower. In order to maintain the spacing of the vertical elements, small cross members are inserted.

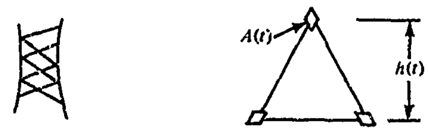


Figure 7-3. Tower Considered

It is assumed that the tower is constructed of a given material with density ρ. Further, it is assumed that βh cubic units of material are required per unit height of tower in order to maintain the spacing of the vertical elements. The coefficient β is to be determined from design experience. For this configuration the total weight of the tower is

$$W = \int_0^L 3\rho\,[A(t) + \beta h(t)]\,dt.$$

Since 3ρ is a constant, W is minimized if and only if

$$V = \int_0^L [A(t) + \beta h(t)]\,dt \qquad (7\text{-}14)$$

is minimum. The objective in the design problem is to choose $A(t)$ and $h(t)$ for $0 < t < L$, so that V is as small as possible and the given conditions are met.

Lateral deflection of the tower due to the lateral wind load is determined by elementary beam theory. The differential equation for displacement is (Ref. 14)

$$EI(t)x'' = -M(t) \tag{7-15}$$

where

E = Young's modulus of the material

$I(t)$ = minimum moment of inertia of the cross-sectional area of the tower

x = lateral displacement

x'' $= \dfrac{d^2x}{dt^2}$

$M(t)$ = bending moment of tower.

The moment of inertia of the cross section is

$$I(t) = \frac{2}{3}A(t)[h(t)]^2. \tag{7-16}$$

In order to prevent damage in handling, it is required that

$$I(t) > I_0 > 0, \quad 0 < t < L.$$

or in the notation of par. 6-4,

$$\phi = I_0 - I(t) < 0, 0 < t < L.$$

If $I(t) = I_0$ gives a tower with angular deflection less than or equal to θ at the top, then this is the optimal tower and no further work is required. On the other hand, if this tower has angular deflection greater than θ, then the tower is not admissible and it is required that the angular displacement is equal to θ. This is the only situation considered here.

In order to fit this problem into the form considered in par. 6-4, define

$$x_1 = x$$

$$x_2 = x_1'$$

with this notation and that of par. 6-4, the second-order equation 7-2 is equivalent to

$$\left.\begin{aligned} x_1' &= x_2 \equiv f_1 \\ x_2' &= -\frac{M}{EI} \equiv f_2. \end{aligned}\right\} \tag{7-17}$$

The design problem will now be solved for two admissible configurations of the tower.

In order to compare results obtained for the various configurations considered, a tower with properties of Table 7-2 will be treated.

TABLE 7-2

CONSTANTS

L = 360 in.	θ = 0.0001 rad
Q = 8.35 lb/in.	β = 0.25 in.
E = 3 x 10^7 lb/in.	I_0 = 172.8 in.^{4}

7-3.2 TOWER WITH ONE DESIGN VARIABLE

For the problems considered in this paragraph, $A(t)$ will be held constant with the value of A. Two ways of mounting the tower on the earth will be chosen. The first method will be to fix the base of the tower rigidly to the earth and leave the top unsupported. The second method will be to pin the lower end to the earth and support the top with guy lines. Since A is constant

$$V = \int_0^L [A + h(t)]\,dt$$

$$= AL + \beta \int_0^L h(t)dt.$$

Minimization of V in this case is equivalent to minimization of

$$J = \int_0^L h(t)dt. \qquad (7\text{-}18)$$

In the notation of par 6-4,

$$f_0 = h(t).$$

7-3.2.1 METHOD 1. TOWER WITH BASE RIGIDLY FASTENED TO THE EARTH

The tower considered here is shown in Fig 7-4. The bending moment $M(t)$ due to the

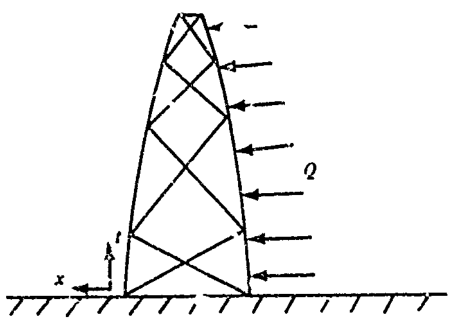


Figure 7-4. Loading of Tower

wind load Q is

$$M(t) = -\frac{Q}{2}(L - t)^2 \qquad (7\text{-}19)$$

where t is measured upward from the bottom of the tower; the other symbols are defined in par 7-3.1. Eqs. 7-17 become

$$\left.\begin{aligned} x_1' &= x_2 &&\equiv f_1 \\ x_2' &= \frac{3Q(L-t)^2}{4EAh^2} &&\equiv f_2 \end{aligned}\right\} \qquad (7\text{-}20)$$

with boundary conditions

$$\left.\begin{aligned} x_1(0) &= 0 \\ x_2(0) &= 0 \\ x_2(L) &= \theta \end{aligned}\right\} \qquad (7\text{-}21)$$

The problem stated here will now be solved using Theorem 6-7. Define

$$H = \lambda_1 x_2 + \lambda_2 \left[\frac{3Q(L-t)^2}{2EAh^2}\right] - \lambda_0 h$$

$$- \mu\left(I_0 - \frac{2}{3}Ah^2\right)$$

and

$$G = \gamma_1 x_1(0) + \gamma_2 x_2(0) + \gamma_3 [x_2(L) - \theta].$$

The conditions of Theorem 6-7 yield

$$\left.\begin{aligned} \lambda_1' &= 0 \\ \lambda_2' &= -\lambda_1 \end{aligned}\right\} \qquad (7\text{-}22)$$

$$0 = \frac{\partial H}{\partial h} = -\lambda_0 - \lambda_2 \left[\frac{6Q(L-t)^2}{2EAh^3}\right]$$

$$+ \mu\frac{4}{3}Ah \qquad (7\text{-}23)$$

$$\lambda_1(L) = 0 \qquad (7\text{-}24)$$

The general solution of Eq. 7-22 is

$$\lambda_1(t) = \xi_1$$

$$\lambda_2(t) = -\xi_2 - \xi_1 t.$$

Condition, Eq. 7-24, implies $\xi_1 = 0$, so

$$\lambda_1(t) = 0$$
$$\lambda_2(t) = -\xi_2. \tag{7-25}$$

For the determination of $h(t)$, two cases must be considered:

Case 1: $\phi = 0$. In this case

$$\frac{2}{3}Ah^2 - I_0 = 0$$

so

$$h = \left(\frac{3I_0}{2A}\right)^{1/2} \equiv h_1$$

Case 2: $\phi < 0$. In this case Eq. 6-128 is

$$\mu\phi = 0$$

and since $\phi \neq 0$, $\mu = 0$. Substituting this result into Eq. 7-23,

$$-\lambda_C + \frac{6\xi_2 Q(L-t)^2}{2EAh^3} = 0. \tag{7-26}$$

If $\lambda_0 = 0$, then, $\xi_2 = 0$ and from Eq. 7-26, $\lambda_1 = \lambda_2 = 0$. This, however, violates the condition stated in the first sentence of Theorem 6-7. Therefore, $\lambda_0 \neq 0$ and it is permissible o choose $\lambda_0 = 2$

Further, since $\lambda_0 > 0$, $\xi_2 > 0$ in order that Eq. 7-26 hold. Eq. 7-26 then yields

$$h = \left[\frac{3\xi_2 Q(L-t)^2}{2EA}\right]^{1/3} \equiv h_2$$

Since Cases 1 and 2 cover all possibilities,

$$h = \left\{ \left[\frac{3\xi_2 Q(L-t)^2}{2EA}\right]^{1/3} \text{ or } \left(\frac{3I_0}{2A}\right)^{1/2} \right\} \tag{7-27}$$

where by Eq. 6-131, the choice in Eq. 7-27 must be made which makes

$$H = -h - \frac{3\xi_2 Q(L-t)^2}{4EAh^2}$$

a maximum.

Let t^* be a point of transition from one expression in Eq. 7-27 to the other. Since Theorem 6-7 requires H to be continuous, when the two expressions of Eq. 7-27 evaluated at t^* are substituted into H, a common value must occur.

Evaluate H as a function of both h_2 and h_1

$$H(h_1) \equiv H_1 = -h_1 - \frac{1}{2}\left[\frac{3\xi_2 Q(L-t)^2}{2EA}\right]\frac{1}{h_1^2}$$

$$= -h_1 - \frac{1}{2}\frac{h_2^3}{h_1^2}$$

$$H(h_2) \equiv H_2 = -h_2 - \frac{1}{2}\left[\frac{3\xi_2 Q(L-t)^2}{2EA}\right]\frac{1}{h_2^2}$$

$$= -h_2 - \frac{1}{2}\frac{h_2^3}{h_2^2}.$$

Then

$$H_2 - H_1 = -\frac{3}{2}h_2 + h_1 + \frac{1}{2}\frac{h_2^3}{h_1^2}$$

$$= \frac{1}{2h_1^2}(2h_1 + h_2)(h_1 - h_2)^2 .$$

If $h_2 < h_1$ at any point $\hat{t}$, then h_2 violates the constraint and $h = h_1$. If $h_2 = h_1$ at any $\hat{t}$, there are no alternative choices for h. If $h_2 > h_1$ at any point t, the choice $h = h_1$ must maximize $H(t)$; this implies $H_2 - H_1 < 0$. But this is impossible from the above because h_1 and h_2 are always non-negative. Therefore, if $h_2 > h_1$, then it is required that $h = h_2$. From this, it is concluded that

$$h(t) = \max\{h_1(t), h_2(t)\} .$$

Since $h(t)$ is defined as the maximum of two continuous functions, h is continuous. It follows that all points t^* of transition from one value of Eq. 7-27 to another can be found by equating the two expressions of $h(t)$.

The point t^* is then determined by

$$h_1(t^*) = \left(\frac{3I_0}{2A}\right)^{1/2} = \left[\frac{3\xi_2 Q(L - t^*)^2}{2EA}\right]^{1/3} = h_2(t^*).$$

This solution yields two values of t^*. The requirement $0 < t^* < L$ results in a unique value of t^*,

$$t^* = L - \left[\frac{3E^2 I_0^3}{2A\xi_2^4 Q^2}\right]^{1/4} . \qquad (7\text{-}28)$$

For $0 < t < L$, the first form of h in Eq. 7-27 is monotone decreasing and is zero at $t = L$. It is, therefore, clear that the second form of h must hold for $t^* < t < L$ and the first form for $0 < t < t^*$. The problem is now to determine ξ_2.

The condition which has not yet been satisfied is $x_2(L) = \theta$. Substituting Eq. 7-27 into Eq. 7-20 yields

$$x_2' = \begin{cases} \dfrac{1}{2}\left(\dfrac{3Q}{2EA}\right)^{1/3} (L - t)^{2/3}\, \xi_2^{-2/3}, \\ \quad \text{for } 0 < t < t^* \\ \dfrac{Q(L - t)^2}{2EI_0}, \quad \text{for } t^* < t < L \end{cases} \qquad (7\text{-}29)$$

with $x_2(0) = 0$.

Integrating Eq. 7-29 first from 0 to t^* (as given by Eq. 7-28) and then from t^* to L yields

$$x_2(L) = \frac{3L^{5/3}}{10}\left(\frac{3Q}{2EA}\right)^{1/3} \xi_2^{-2/3} - \frac{E^{1/2} I_0^{5/12}(2^{3/4} \times 3 + 10 I_0^{5/6})}{(24)^{1/4}\, 10 Q^{1/2} A^{3/4}}\,\xi^{-3/2} = \theta. \qquad (7\text{-}30)$$

Eq. 7-30 is solved numerically for ξ_2. Once ξ_2 is determined, then Eqs. 7-27 and 7-28 completely specify the tower. Results are shown in Table 7-3(A) and Fig. 7-5(A).

This design problem has been solved analytically. As will become apparent as more realistic r s are treated, one should not expect ain solutions in this way. In most problems, numerical methods must be applied to solve the differential equations arising in the Theorem 6-7

$$\lambda_1(t) = \xi_1$$

$$\lambda_2(t) = -\xi_2 - \xi_1 t.$$

Condition, Eq. 7-24, implies $\xi_1 = 0$, so

$$\lambda_1(t) = 0$$
$$\lambda_2(t) = -\xi_2. \qquad (7\text{-}25)$$

For the determination of $h(t)$, two cases must be considered:

Case 1: $\phi = 0$. In this case

$$\frac{2}{3}Ah^2 - I_0 = 0$$

so

$$h = \left(\frac{3I_0}{2A}\right)^{1/2} \equiv h_1$$

Case 2: $\phi < 0$. In this case Eq. 6-128 is

$$\mu\phi = 0$$

and since $\phi \neq 0$, $\mu = 0$. Substituting this result into Eq. 7-23,

$$-\lambda_0 + \frac{6\xi_2 Q(L-t)^2}{2EAh^3} = 0. \qquad (7\text{-}26)$$

If $\lambda_0 = 0$, then, $\xi_2 = 0$ and from Eq. 7-26, $\lambda_1 = \lambda_2 = 0$. This, however, violates the condition stated in the first sentence of Theorem 6-7. Therefore, $\lambda_0 \neq 0$ and it is permissible to choose $\lambda_0 = 2$.

Further, since $\lambda_0 > 0$, $\xi_2 > 0$ in order that Eq. 7-26 hold. Eq. 7-26 then yields

$$h = \left[\frac{3\xi_2 Q(L-t)^2}{2EA}\right]^{1/3} \equiv h_2$$

Since Cases 1 and 2 cover all possibilities,

$$h = \left\{ \left[\frac{3\xi_2 Q(L-t)^2}{2EA}\right]^{1/3} \text{ or } \left(\frac{3I_0}{2A}\right)^{1/2} \right\} \qquad (7\text{-}27)$$

where by Eq. 6-131, the choice in Eq. 7-27 must be made which makes

$$H = -h - \frac{3\xi_2 Q(L-t)^2}{4EAh^2}$$

a maximum.

Let t^* be a point of transition from one expression in Eq. 7-27 to the other. Since Theorem 6-7 requires H to be continuous, when the two expressions of Eq. 7-27 evaluated at t^* are substituted into H, a common value must occur.

Evaluate H as a function of both h_2 and h_1.

$$H(h_1) \equiv H_1 = -h_1 - \frac{1}{2}\left[\frac{3\xi_2 Q(L-t)^2}{2EA}\right]\frac{1}{h_1^2} = -h_1 - \frac{1}{2}\frac{h_2^3}{h_1^2}$$

$$H(h_2) \equiv H_2 = -h_2 - \frac{1}{2}\left[\frac{3\xi_2 Q(L-t)^2}{2EA}\right]\frac{1}{h_2^2} = -h_2 - \frac{1}{2}\frac{h_2^3}{h_2^2}.$$

TABLE 7-3(A)
WEIGHTS OF SIMPLY SUPPORTED TOWERS, ONE DESIGN VARIABLE

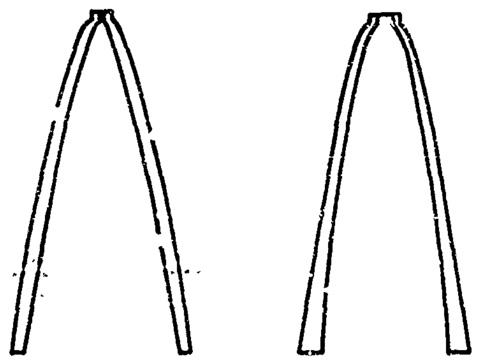
A = 6.0 in.^{2}	W = 2121.5 lb
A = 6.6 in.^{2}	W = 2112.4 lb
A = 6.9 in.^{2}	W = 2111.4 lb
A = 7.2 in.^{2}	W = 2112.3 lb
A = 7.9 in.^{2}	W = 2119.5 lb

TABLE 7-3(B)
WEIGHTS OF GUY-LINE SUPPORTED TOWERS, ONE DESIGN VARIABLE

A = 3.9 in.^{2}	W = 1362.1 lb
A = 4.2 in.^{2}	W = 1357.6 lb
A = 4.434 in.^{2}	W = 1356.6 lb
A = 4.5 in.^{2}	W = 1356.7 lb
A = 4.8 in.^{2}	W = 1358.8 lb

TABLE 7-3(C)
WEIGHTS OF TOWERS

	Cantilevered	Cantilevered	Cantilevered	Guy-line Supported	Guy-line Supported	Guy-line Supported
Number of Design Variables	0	1	2	0	1	2
Best Weight	W = 2440.6 lb	W = 2111.4 lb	W = 1827.9 lb	W = 1563.99 lb	W = 1356.6 lb	W = 1265.71 lb
	h = 63.7 in.	h_{max} = 91.4 in.	h_{max} = 80.2 in.	h = 46 in.	h_{max} = 46.5 in.	h_{max} = 36.55 in.
	A = 7.96 in.^{2}	A = 6.97 in.^{2}	A_{max} = 10.03 in.^{2}	A = 3.84 in.^{2}	A = 4.434 in.^{2}	A_{max} = 4.95 in.^{2}



(A) One Control Variable (B) Two Control Variables

Figure 7-5. Tower With Base Rigidly Fastened to the Earth

7-3.2.2 METHOD 2. TOWER WITH BASE PINNED TO EARTH AND WITH TOP SUPPORTED BY GUY LINES

The tower considered here is shown in Fig. 7-6. It is convenient here to locate the coordinate system at the top of the tower. The bending moment generated by the uniform wind load Q is $M = -Q/2\ t(t-L)$ so the differential equation for bending is

$$EI(h)x'' = \frac{Q}{2}\ t(t-L)\ .$$

Define $x_1 = x$ and $x_2 = x_1'$; this is equivalent to

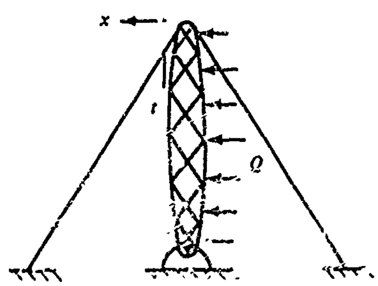


Figure 7-6. Tower With Guy Lines

$$\left.\begin{aligned} x_1' &= x_2 \equiv f_1 \\ x_2' &= \frac{Qt(t-L)}{2EI} \equiv f_2 \end{aligned}\right\} \quad (7\text{-}31)$$

The boundary conditions in this case are

$$\left.\begin{aligned} x_1(0) &= 0 \\ x_2(0) &= \theta \\ x_1(L) &= 0 \end{aligned}\right\} \quad (7\text{-}32)$$

The quantity to be minimized is still given by Eq. 6-17. In the problem considered here,

$$H = \lambda_1 x_2 + \lambda_2 \left[\frac{3Qt(t-L)}{4EAh^2}\right]$$

$$-\lambda_0 h - \mu(I_0 - \frac{2}{3}Ah^2)$$

$$G = \gamma_1 x_1(0) + \gamma_2[x_2(0) - \theta] + \gamma_3 x_1(L)\ .$$

Conditions of Theorem 6-7 are

$$\frac{d\lambda_1}{dt} = \frac{\partial H}{\partial x} = 0$$

$$\frac{d\lambda_2}{dt} = \frac{\partial H}{\partial x_2} = -\lambda_1$$

so

$$\lambda_1(t) = \xi_1$$

$$\lambda_2(t) = -\xi_2 - \xi_1 t\ .$$

Also,

$$\lambda_2(L) = 0$$

This equation implies $\xi_2 = -\xi_1 L$, so

$$\lambda_1 = \xi_1$$

$\lambda_2 = \xi_1(L - t)$.

Two cases must now be considered:

Case 1. $\phi = 0$. In this case

$$\frac{2}{3}Ah^2 - I_0 = 0$$

so

$$h_1 = \left(\frac{3I_0}{2A}\right)^{1/2} . \tag{7-33}$$

Case 2. $\phi < 0$. In this case $\mu = 0$. Substituting into $\partial H/\partial h = 0$,

$$-\lambda_0 - \frac{\xi_1(L-t)\,6Qt(t-L)}{2EAh^3} = 0 .$$

As in the previous case $\lambda_0 \neq 0$ so it is permissible to put $\lambda_0 = 2$ and obtain

$$h_2 = \left[\frac{3\xi_1 Qt(L-t)^2}{2EA}\right]^{1/3} . \tag{7-34}$$

Eqs. 7-33 and 7-34 along with Eq. 6-131 yield

$$h(t) = \left\{ \left(\frac{3I_0}{2A}\right)^{1/2} \quad \text{or} \quad \left[\frac{3\xi_1 Qt(L-t)^2}{2EA}\right]^{1/3} \right\} \tag{7-35}$$

the choice in Eq. 7-35 being made which makes

$$H = -h - \frac{3\xi_1 Qt(L-t)^2}{4EAh^2}$$

largest.

The problem is thus solved when ξ_1 is determined. In this case, analytical integration of Eq. 7-31 and solution of the boundary conditions, Eq. 7-32, for ξ_1 are not feasible. Therefore, the shooting technique of par. 6-5 is employed. Numerical results for this problem are given in Fig. 7-7(A) and Table 7-3(B).

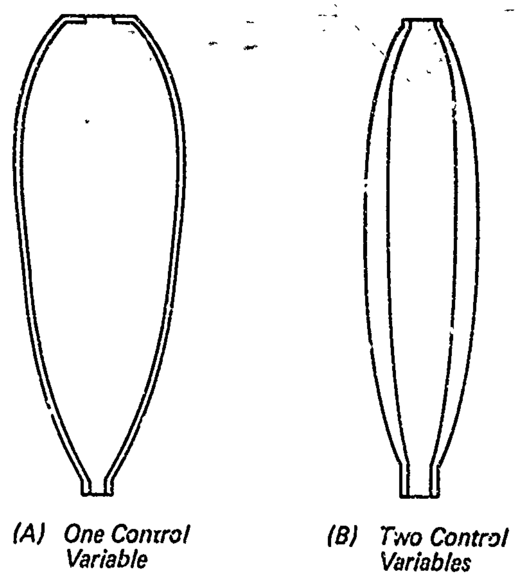


Figure 7-7. Tower With Base Simply Supported and Top Supported With Guy Lines

The modification of Newton's Algorithm consists of using a correction of at most 10% of the values of the unknown iteration parameters. It has been found in particular problems that where the Newton Method fails to converge, this method will converge. The rate of convergence, particularly for the first few iterations, is slowed by the modification, however.

7-3.3 TOWER WITH TWO DESIGN VARIABLES

The same two methods of supporting the towers will again be considered. Here, however, both A and h will be allowed to vary along the tower and play the role of design variables.

Eqs. 7-16 and 7-17 remain the same in the present problem. However, from Eq. 7-14

$$f_0 = A(t) + \beta h(t) .$$

7-3.3.1 METHOD 1. TOWER WITH BASE RIGIDLY FASTENED TO THE EARTH

Fig. 7-4 applies and Eqs. 7-19, 7-20, and 7-21 hold in this problem. Substitution into Eq. 6-108 yields

$$H = \lambda_1 x_2 + \lambda_2 \left[\frac{3Q(L-t)^2}{4EAh^2}\right] - \lambda_0 (A + \beta h) - \mu\left(I_0 - \frac{2}{3} Ah^2\right) .$$

The equations for λ_1 and λ_2 are just as in the preceding work, so again,

$$\lambda_1(t) = 0$$

$$\lambda_2(t) = -\xi_2 .$$

Since h plays the role of u_1 and A the role of u_2, Eq. 6-122 is:

$$\frac{\partial H}{\partial h} = -\lambda_0 \beta + \frac{\xi_2 3Q(L-t)^2}{2EAh^3} + \mu \frac{4}{3} Ah = 0$$

$$\frac{\partial H}{\partial A} = -\lambda_0 + \frac{\xi_2 3Q(L-t)^2}{4EAh} + \mu \frac{2}{3} h^2 = 0. \tag{7-36}$$

As before, two cases must be considered:

Case 1. $\phi < 0$. This implies $\mu = 0$. From Eq. 7-36, it is clear that $\lambda_0 = 0$ implies $\xi_2 = 0$ which contradicts Theorem 6-7. Therefore, it is permissible to take $\lambda_0 = 1$. The system is then two equations for h and A with solution

$$\left.\begin{aligned} h(t) &= \left(\frac{3\xi_2 Q}{\beta^2 E}\right)^{1/4} (L-t)^{1/2} \\ A(t) &= \left(\frac{3\xi_2 Q \beta^2}{E}\right)^{1/3} \frac{(L-t)^{1/2}}{2} . \end{aligned}\right\} \tag{7-37}$$

Case 2. $\phi = 0$. This is

$$\frac{2}{3} Ah^2 = I_0 \tag{7-38}$$

Eq. 7-38 along with Eq. 7-36 is a system of three equations in h, A, and μ and solving for h and A yields

$$\left.\begin{aligned} h(t) &= \left(\frac{3I_0}{4\beta}\right)^{1/3} \\ A(t) &= \left(\frac{3I_0\beta^2}{2}\right)^{1/3} . \end{aligned}\right\} \tag{7-39}$$

The design variables are chosen by Eqs. 7-37 and 7-39, depending on which makes H largest. The problem with differential equations, Eq. 7-20, and boundary conditions, Eq. 7-21, is now treated by the shooting technique of par. 6-5. Numerical results are given in Table 7-3(C).

7-3.3.2 METHOD 2. TOWER WITH BASE PINNED TO EARTH AND WITH TOP SUPPORTED BY GUY LINES

Fig. 7-6 applies and Eqs. 7-31 and 7-32 hold for this problem. Substituting into Eq. 6-108 yeilds

$$H = \lambda_1 x_2 + \lambda_2 \left[\frac{3Qt(t-L)}{4EAh^2}\right] - \lambda_0 (A + \beta h) - \mu\left(I_0 - \frac{2}{3} Ah^2\right) .$$

The equations for λ_1 and λ_2 are just as in the preceding case, so again

$$\lambda_1 = \xi_1$$

$$\lambda_2 = \xi_1(L - t)$$

Eqs. 6-122 for the design variables are

$$\left.\begin{aligned} \frac{\partial H}{\partial h} &= -\lambda_0\beta + \frac{\xi_1 3Qt(L-t)^2}{2EAh^2} \\ &\quad + \mu\frac{4}{3}Ah = 0 \\ \frac{\partial H}{\partial A} &= -\lambda_0 + \frac{\xi_1 3Qt(L-t)^2}{4EA^2h^2} \\ &\quad + \mu\cdot\frac{2}{3}h^2 = 0\,. \end{aligned}\right\} \quad (7\text{-}40)$$

As before, two cases must be considered:

Case 1. $\phi < 0$. This implies $\mu = 0$.

From Eq. 7-40, it is clear that if $\lambda_0 = 0$, then $\xi_1 = 0$, and $\lambda_1 = \lambda_2 = 0$. This contradicts the theorem, so $\lambda_0 \neq 0$ and it is permissible to take $\lambda_0 = 1$. The system is then a set of two equations for h and A which yields

$$\left.\begin{aligned} h(t) &= \left(\frac{3\xi_1 Q}{\beta^2 E}\right)^{1/4} t^{1/4}(L-t)^{1/2} \\ A(t) &= \frac{1}{2}\left(\frac{3\xi_1 Q\beta^2}{E}\right)^{1/4} t^{1/4}(L-t)^{1/2} \end{aligned}\right\} \quad (7\text{-}41)$$

Case 2. $\phi = 0$. This is

$$\frac{2}{3}Ah^2 = I_0. \quad (7\text{-}42)$$

Eq. 7-42 along with Eq. 7-40 is a system of three equations for h, A, and μ. Eliminating μ and solving for h and A yields

$$\left.\begin{aligned} h(t) &= \left(\frac{3I_0}{4\beta}\right)^{1/3} \\ A(t) &= \left(\frac{3I_0\beta^2}{2}\right)^{1/3} \end{aligned}\right\} \quad (7\text{-}43)$$

The design variables are chosen as in Eqs. 7-41 or 7-43, depending on which makes H largest. The problem with differential equations and boundary conditions, Eq. 7-20, is now solved by the shooting technique. Numerical results are given in Table 7-3(C).

7-3.4 DISCUSSION OF RESULTS

For both types of tower considered (simply supported and towers with top supported by guy lines), the variables A and h could be fixed at constant values, large enough that deflection requirements are met. For a given configuration of the tower, there is one pair of constant values A and h which yield a tower at least as light as any other combination of constant A and h. For both types of tower, finding these values is a matter of simple algebra. Results for both towers are given in Table 7-3(C), referred to as towers with no design variables. With A held constant and $h(t)$ treated as a design variable, the problem can be solved for several different values of A for both types of tower. Summaries of these solutions are given in Table 7-3(A) and Table 7-3(B). If the tower weights are then plotted as a function of fixed values of A, a minimum, or best, weight can be found (Table 7-3(C)). These optimum towers show a reduction in weight over the no design variable case of 13.5% and 13% for the simply supported and guy-line supported towers, respectively. Results for solutions of the problem when both $A(t)$ and $h(t)$ are treated as design variables are also given in Table 7-3(C). These represent reductions in weight, over the case where only the spacing $h(t)$ is allowed to vary, of 6.1% and 6.7% for the simply supported and guy-line supported towers, respectively. Similarly, these towers represent respective reductions in weight over the completely uniform (no design variables) tower of 18.8% and 19%.

All four configurations of the structure described in the preceding subparagraphs have been successfully prescribed by a digital computer approach. An IBM 360-65 Computer was used and the programs employed Runge-Kutta integration with the Newton's Method described in the text. Convergence depended on getting a good starting value for the multiplier ξ. This was made more difficult by the fact that ξ has no physical significance so that, consequently, engineering intuition was no help. To find a sufficiently close starting value for ξ, (i.e., a value for which the error is of order two or less) several different values of ξ were investigated, each increasing from the previous value by a factor of 10, and the first value very close to zero. Once a starting value that would allow Newton's Method iterate effectively was found, convergence occurred in ten or less iterations, taking less than two minutes of computer time. This time could be reduced with increased sophistication of the computer program.

Other optimal design problems can be approached with the method of this paragraph. Also different parameters could be treated as variables. For instance, several materials of different densities and stiffness characteristics could be used in the same structure, and the choice could be left as a design parameter. The loading could be a function of the height above the ground rather than be constant. Other restrictions might also be imposed, maximum width or maximum deflection at a given point could be among these.

Note that Figs. 7-5 and 7-6 are not scale drawings of the towers, but are representative of the general shape of the respective towers, as viewed on one face. Sample profiles of the four possible structures are presented in Table 7-3(A), (B), and (C).

7-4 MINIMUM WEIGHT DESIGN OF BEAMS WITH INEQUALITY CONSTRAINTS ON STRESS AND DEFLECTION

The problems treated thus far in this chapter have only design variable inequality constraints. In engineering design one often encounters problems in which it is required that the state of the system satisfies inequality constraints. As seen in Chapter 6, state variable constraints are more tedious to treat and have features not encountered in problems without state constraints. A class of beam design problems including state variable constraints is presented in this paragraph to illustrate some of the features and difficulties that can arise in this difficult class of problems. While the problems solved are of limited practical value, they do illustrate typical features that can arise in state variable constrained problems

7-4.1 STATEMENT OF THE PROBLEM

Beams which are loaded in a general way (such as in Fig. 7-8) are considered in this paragraph. The cross sections of the beams are assumed to depend on a vector parameter $u(t) = [u_1(t), u_2(t), ..., u_m(t)]$ and to be symmetric with respect to vertical and horizontal axes. The vertical axis of symmetry is assumed to lie in the plane of loading. The beams are made of a homogeneous, isotropic,

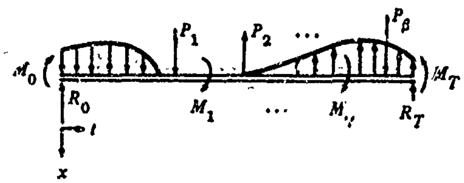


Figure 7-8. Beam Loaded in a General Way

linearly elastic material with Young's modulus E. Small deflection, elementary beam theory is used throughout this paragraph. Also, the effect of the weight of the beam on deflection is neglected.

Since, for a particular beam, the cross section is determined by $u(t)$,

$A(t) = A[u(t)]$ (cross-sectional area), (7-44)

$I(t) = I[u(t)]$ (moment of inertia), (7-45)

$b(t,d) = b[u(t),d]$ (width of cross section at d), (7-46)

$Q(t,d) = Q[u(t),d]$ (first moment of area above d, about the neutral axis), (7-47)

and

$C(t) = C[u(t)]$ (half-depth of beam), (7-48)

where d is the distance above the neutral axis of the cross section.

For each t, let $d_i\,[t,u(t),M(t),V(t)]$, $i = 1$ and 2, be the distances from the neutral axis where

$$\sigma_p(t) = \frac{1}{2}\left[\frac{|M(t)|\,d_1(t)}{I(t)}\right] + \frac{1}{2}\left\{\frac{M^2(t)d_1^2(t)}{I^2(t)} + 4\left(\frac{V^2(t)Q^2[t,d_1(t)]}{I^2(t)b^2[t,d_1(t)]}\right)\right\}^{1/2}$$

and

$$\tau_p(t) = \frac{1}{2}\left\{\frac{M^2(t)d_2^2(t)}{I^2(t)} + \left(\frac{V^2(t)Q^2[t,d_2(t)]}{I^2(t)b^2[t,d_2(t)]}\right)\right\}^{1/2}$$

respectively, are the maximum principal stresses that occur in the cross section at t. The values d_1 and d_2 may be determined by the methods of ordinary calculus, for each t.

The problem is to determine $u(t)$ so that the beam, subjected to a given loading, contains as little material as possible and still satisfies the following conditions:

1. Principal normal stress is less than or equal to some allowable normal stress σ_{max}.
2. Principal shear stress is less than or equal to some allowable shear stress τ_{max}.
3. Stiffness is bounded away from zero (otherwise an infinitesimal change in

load can cause the deflection to be discontinuous).

4. Beam deflection at each point is bounded by two given functions $X_1(t)$ and $X_2(t)$, i.e., $X_1(t) < x(t) < X_2(t)$ with $X_1(t) < X_2(t)$.

In structural design problems, it is frequently sufficient to require only that the maximum flexural stress be less than σ_{max} and the maximum direct shear stress be less than τ_{max}. These conditions are considerably easier to enforce than the conditions on maximum principal stress.

If the beam is subjected to several loadings, then the problem is more tedious but is no more difficult mathematically. Corresponding to each loading there is a a deflection curve, bending stress, and shear stress that must satisfy the stated conditions.

Further, since the beams are made of homogeneous material, the weight of a beam will be minimum if and only if its volume is minimum. Therefore, in the following the quantity to be minimized will be volume.

The given problem is now stated mathematically: A vector function $u(t)$ is sought which causes the functional

$$J = \int_0^T A\{u(t)\}\,dt \tag{7-49}$$

to be a minimum subject to the following conditions:

$$\frac{d^2}{dt^2}\left\{EI[u(t)]\,\frac{d^2x}{dt^2}\right\} = q(t) \tag{7-50}$$

at all but a finite number of points in $(0, T)$, where $q(t)$ is distributed load;

$$g_s[x^{(i)}(0), x^{(i)}(T)] = 0,\ s = 1, ..., B,\ i < 3; \tag{7-51}$$

$$\sigma_p(t) < \sigma_{max} \tag{7-52}$$

for all t in $(0,T)$, where $M(t)$ is bending moment;

$$\tau_p(t) < \tau_{max} \tag{7-53}$$

for all t in $(0,T)$ where $V(t)$ is shear;

$$I[u(t)] > I_0\ , \tag{7-54}$$

where I_0 is a constant greater than zero; and for all t in $(0,T)$,

$$X_1(t) < x(t) < X_2(t)\ . \tag{7-55}$$

It is assumed that the functions appearing above have the following properties:

1. q has a piecewise continuous derivative in $(0,T)$.

2. A, I, C, and Q are piecewise twice continuously differentiable.

3. X_1 and X_2 have continuous second derivatives in $(0,T)$.

A solution is sought with the following properties:

1. $u_j(t)$, $j = 1, ..., m$, are piecewise continuous in $(0,T)$.

2. $x(t)$ is piecewise four times continuously differentiable in $(0,T)$.

In case only maximum bending stress and

maximum direct shear stress are to be bounded by σ_{max} and τ_{max}, respectively, constraints, Eqs. 7-52 and 7-53, are replaced by

$$|\sigma(t)| = \frac{|M(t)|\, C[u(t)]}{I[u(t)]} < \sigma_{max} \qquad (7\text{-}56)$$

and

$$|\tau(t)| = \frac{|V(t)|\, Q\,[u(t), d_3(t)]}{I[u(t)]\, b\,[u(t), d_3(t)]} < \tau_{max} \qquad (7\text{-}57)$$

for all t in $(0,T)$ where $d_3(t) = d_3\,[t, u(t), M(t), V(t)]$ is the distance from the neutral axis where the absolute value of the direct shear stress is largest. The distance $d_3(t)$ may be determined by the methods of ordinary calculus.

In the case of beam design with multiple loading requirements, there is still just one design variable $u(t)$. However, corresponding to each loading there is an additional state variable (deflection) that must satisfy conditions identical in form to Eqs. 7-50 through 7-55. The problem is to determine $u(t)$ so that the functional, Eq. 7-49, is minimum subject to the condition that Eqs. 7-50 through 7-55 are satisfied for each loading.

7-4.2 NECESSARY CONDITIONS FOR THE BEAM DESIGN PROBLEM

The treatment which follows applies only to statically determinate beams, i.e., beams loaded in such a way that reactions at all supports (and hence also shear and bending moment) are determined completely by the conditions for equilibrium of the beam. Changes in formulation of the problem which are necessary to consider statically indeterminate problems will be indicated below.

In the statically determinate case, the differential equation, Eq. 7-50, and the boundary conditions, Eq. 7-51, reduce to

$$\frac{d^2x}{dt^2} = -\frac{M(t)}{EI(u)} \qquad (7\text{-}58)$$

and

$$g_s\,[x(0), x'(0), x(T), x'(t)] = 0,$$

$$s = 1, 2. \qquad (7\text{-}59)$$

The boundary-value problem, Eqs. 7-58 and 7-59, is equivalent to a boundary-value problem with a system of first-order equations. The new problem may be written as

$$\frac{dx_1}{dt} = x_2 \qquad (7\text{-}60)$$

$$\frac{dx_2}{dt} = -\frac{M(t)}{EI(u)}, \qquad (7\text{-}61)$$

and

$$g_s\,[x_1(0), x_2(0), x_1(T), x_2(T)] = 0,$$

$$s = 1, 2 \qquad (7\text{-}62)$$

where x_1 is defined to be x. In terms of this notation, Eq. 7-57 becomes

$$X_1(t) < x_1(t) < X_2(t).$$

It is now clear that the beam design problem is contained in the class of problems to which par. 6-4 applies. The quantities appearing in par. 6-4 will now be identified with the physical quantities associated with beam so that necessary conditions for the beam design problem may be stated.

Conditions, Eqs. 7-52, 7-53, and 7-54,

correspond to Eqs. 6-101; Eq. 7-55 corresponds to two restrictions of the type expressed by Eq. 6-132. The differential equations, Eqs. 7-60 and 7-61, correspond to Eqs. 6-97, where

$$f_1 = x_2 \tag{7-63}$$

$$f_2 = -\frac{M(t)}{EI(u)} \tag{7-64}$$

and in Eq. 6-96

$$f_0 = A(u). \tag{7-65}$$

The ends of the beam are located at the known points $t = 0$ and $t = T$. Therefore, t^0 and t^1 in the general variational problem are known. Also, boundary conditions will generally be separated; i.e., some conditions will be given at 0 and others at T.

The state variable constraints, Eq. 7-55, are of second-order since

$$\frac{d}{dt}(X_2 - x_1) = X_2' - x_1' = X_2' - x_2$$

is not an explicit function of u, but

$$\frac{d^2}{dt^2}(X_2 - x_1) = X_2''(t) - x_2'(t)$$

$$= X_2''(t) + \frac{M(t)}{EI(u)}$$

is an explicit function of u. The same argument holds for $x_1 - X_1 > 0$.

In terms of the notation of Eq. 6-101,

$$\phi_1 = \sigma_p(t) - \sigma_{max} < 0$$

$$\phi_2 = \tau_p(t) - \tau_{max} < 0$$

$$\phi_3 = I_0 - I(u) < 0$$

$$\phi_4 = y_1(t) - Y_2(t) < 0$$

$$\phi_5 = Y_1(t) - y_1(t) < 0$$

$$\phi_{4,2} = \begin{cases} \phi_4, & \phi_4 < 0 \\ -\dfrac{M(t)}{EI(u)} - X_2''(t), & \phi_4 = 0 \end{cases} \tag{7-66}$$

and

$$\phi_{5,2} = \begin{cases} \phi_5, & \phi_5 < 0 \\ \dfrac{M(t)}{EI(u)} + X_1''(t), & \phi_5 = 0 \end{cases} \tag{7-67}$$

Since the only explicit dependence of Eqs. 7-52, 7-53, 7-54, 7-63, 7-64, and 7-65 on t is through $M(t)$ and $V(t)$, the points of discontinuity of functions (t^j in par. 6-4) are denoted ω_α, including points of discontinuity of $M(t)$ or $V(t)$; i.e., the ω_α correspond to points of application of concentrated loads or moments. Therefore, the ω_α are known.

At each point $t_r^{\sim}$, the deflection curve is tangent to one of the curves

$$x_1 = X_1(t) \tag{7-68}$$

or

$$x_1 = X_2(t) \tag{7-69}$$

and there is a neighborhood of $t_r^{\sim}$ in which this is the only point of tangency.

At each t_δ^-, the deflection curve becomes tangent to one of the curves, Eqs. 7-68 or 7-69. Further, there is a neighborhood of t_δ^- in which Eq. 7-55 is a strict inequality to the left of t_δ^-, and Eq. 7-68 or Eq. 7-69 holds to the right of t_δ^-. At the corresponding point t_δ^+, the deflection curve leaves Eq. 7-68 or Eq. 7-69. Eq. 7-68 or Eq. 7-69 then holds in (t_δ^-, t_δ^+), and Eq. 7-55 is again a strict inequality immediately to the right of t_δ^+.

If $x_1(0)$ or $x_1(T)$ is not fixed by the boundary conditions of Eq. 7-62, then one part of Eq. 7-55 may be an equality at 0 or T. In this case ϕ_5'' of Eqs. 7-66 and 7-67 need not be zero.

The t_η^* are points where one or more of Eqs. 7-52, 7-53, and 7-54 changes from strict inequality to equality. Condition, Eq. 6-105, is assumed to hold at these points.

According to Eqs. 6-106, 6-108

$$\left.\begin{aligned}
\bar{H} &= -\lambda_0 A(u) + \lambda_1 x_2 - \lambda_2\left[\frac{M(t)}{EI(u)}\right] \\
&\quad - \mu_1\phi_1 - \mu_2\phi_2 - \mu_3\phi_3 \\
&\quad - \mu_4\phi_{4,2} - \mu_5\phi_{5,2} \\
\bar{G} &= \sum_{s=1}^{B} \gamma_s g_s + \sum_{\alpha} \gamma_{B+\alpha}(t - \omega_\alpha) \\
G &= \tau_{0,4,\delta}^- \phi_4(t_\delta^-) + \tau_{1,4,\delta}^- \phi_4'(t_\delta^-) \\
&\quad + \tau_{0,5,\delta}^- \phi_5(t_\delta^-) + \tau_{1,5,\delta}^- \phi_5'(t_\delta^-) \\
&\quad + \tau_{0,4,r}^- \phi_4(t_r^-) + \tau_{1,4,r}^- \phi_4'(t_r^-) \\
&\quad + \tau_{0,5,r}^- \phi_5(t_r^-) + \tau_{1,5,r}^- \phi_5'(t_r^-)
\end{aligned}\right\} \quad (7\text{-}70)$$

Theorem 6-9 and Eq. 7-70 yield Theorem 7-1.

Theorem 7-1 Necessary conditions for the minimum weight beam problem are

$$\frac{dx_1}{dt} = x_2 \quad (7\text{-}71)$$

and

$$\frac{dx_2}{dt} = -\frac{M(t)}{EI(u)} \quad (7\text{-}72)$$

at all but a finite number of points in $(0,T)$;

$$\lambda_0 \geq 0$$

$$\lambda_1 = \xi_1$$

$$\lambda_2 = \xi_2 - \xi_1 t$$

and

$$\lambda_0^2 + \lambda_1^2 + \lambda_2^2 > 0$$

for all t in $(0,T)$, where ξ_1 and ξ_2 may have different values in subintervals of $(0,T)$ which are bounded by the t_r and t_δ^-;

$$\left.\begin{aligned}
&-\lambda_0\frac{\partial A(u)}{\partial u_j} - \lambda_2\frac{\partial}{\partial u_j}\left[\frac{M(t)}{EI(u)}\right] \\
&-\frac{\partial}{\partial u_j}[\mu_1\phi_1 + \mu_2\phi_2 + \mu_3\phi_3 \\
&+ \mu_4\phi_{4,2} + \mu_5\phi_{5,2}] = 0,\ j = 1, \ldots, m,
\end{aligned}\right\} \quad (7\text{-}73)$$

$$\mu_i\phi_i = 0,\ i = 1, 2, 3, \quad (7\text{-}74)$$

and

$$\mu_4\phi_{4,2} = \mu_5\phi_{5,2} = 0 \quad (7\text{-}75)$$

at all but a finite number of points in $(0,T)$,

$$\lambda_i(\omega_\alpha + 0) - \lambda_i(\omega_\alpha - 0) = 0,$$

$$i = 1, 2 \quad (7\text{-}76)$$

$$\lambda_1(t_r^- + 0) - \lambda_1(t_r^- - 0) - \tau_{04r}^-$$
$$+ \tau_{05r}^- = 0 \qquad (7\text{-}77)$$

$$\lambda_2(t_r^- + 0) - \lambda_2(t_r^- - 0) - \tau_{14r}^-$$
$$+ \tau_{15r}^- = 0 \qquad (7\text{-}78)$$

$$\lambda_1(t_\delta^- + 0) - \lambda_1(t_\delta^- - 0) - \tau_{04\delta}^-$$
$$+ \tau_{05\delta}^- = 0 \qquad (7\text{-}79)$$

$$\lambda_2(t_\delta^- + 0) - \lambda_2(t_\delta^- - 0) - \tau_{14\delta}^-$$
$$+ \tau_{15\delta}^- = 0 \qquad (7\text{-}80)$$

$$\bar{H}(\delta_x + 0) - \bar{H}(\omega_\alpha - 0) + \gamma_{B+\alpha} = 0$$

$$\bar{H}(t_r^- + 0) - \bar{H}(t_r - 0) - \tau_{04}^- X_2'(t_r^-)$$
$$- \tau_{14}^- X_2''(t_r^-) + \tau_{05r}^- X_1'(t_r^-)$$
$$+ \tau_{15r}^- X_1''(t_r^-) = 0 \qquad (7\text{-}81)$$

$$\bar{H}(t_\delta^- + 0) - \bar{H}(t_\delta^- - 0) - \tau_{04\delta}^- X_2'(t_\delta^-)$$
$$- \tau_{14\delta}^- X_2''(t_\delta^-) + \tau_{05\delta}^- X_1'(t_\delta^-)$$
$$+ \tau_{15\delta}^- X_1''(t_\delta^-) = 0 \qquad (7\text{-}82)$$

$$\bar{H}(t_\eta^* + 0) - \bar{H}(t_\eta^* - 0) = 0$$

and

$$\bar{H}(t_\delta^+ + 0) - \bar{H}(t_\delta^+ - 0) = 0$$

for all α, r, δ, and η; at each of the points $S = t_r^-$ and t_δ^-, either

$$\phi_1(S) = \phi_1'(S) = 0 \qquad (7\text{-}83)$$

or

$$\phi_2(S) = \phi_2'(S) = 0, \qquad (7\text{-}84)$$

the boundary conditions $g_s = 0$, $s = 1, 2$, must be satisfied along with the conditions

$$\lambda_i(0) = \sum_{s=1}^{2} \gamma_s \frac{\partial g_s}{\partial x_i(0)}, \; i = 1, 2 \qquad (7\text{-}85)$$

and

$$\lambda_i(T) = \sum_{s=1}^{2} \gamma_s \frac{\partial g_s}{\partial x_i(T)}, \; i = 1, 2; \qquad (7\text{-}86)$$

and the Weierstrass condition

$$\bar{H}(x_i, U, \lambda_i, t) < \bar{H}(\lambda_i, u, \lambda_i, t)$$

must be satisfied for each t in $(0,T)$, where U is any function which along with x_1 and x_2 satisfies Eqs. 7-52, 7-53, 7-54, 7-55, 7-60, 7-61, and 7-62 with u replaced by U. The statement of Theorem 7-1 is now complete.

If there is only a scalar control variable $u(t)$, then the condition of Eq. 6-105 will be violated at points t_η^* which are intersections of intervals in which Eq. 7-52, 7-53, or 7-54 is an equality. With an additional hypothesis, however, the conclusions of Theorem 7-1 are still valid.

At a point $t^* \neq \omega_\alpha$, it is assumed that $\psi_1 = \psi_2 = 0$, where $\psi_1 > 0$ and $\psi_2 > 0$ are any two of the constraints of Eqs. 7-52, 7-53, and 7-54. If $\hat{\psi}$ is defined as $\hat{\psi} = \min(\psi_1, \psi_2)$, then $\hat{\psi} > 0$ replaces the conditions $\psi_1 > 0$ and $\psi_2 > 0$. It is assumed that $\partial\psi_1/\partial u$ and $\partial\psi_2/\partial u$ are not zero at t^* The new constraint now satisfies the conditions of Eq. 6-105.

If Theorem 6-8 is applied to the new formulation of the problem, the result is identical to Theorem 7-1 with the exception of Eqs. 7-73, 7-74, and 7-76. However, the new conditions on u are identical to those implied by Eqs. 7-73, 7-74, and 7-76. The roles played by μ_1 and μ_2 in Theorem 7-1 would simply be combined in a new variable $\hat{\mu}$. This result may be stated as Corollary 7-1.

Corollary 7-1: Let there be a scalar design variable $u(t)$ and assume that any two of the inequality constraints, Eqs. 7-52, 7-53, and 7-54 are equalities at t^*. If the first partial derivatives of these two constraint functions with respect to u are not zero at t^*, then Theorem 7-1 holds.

One further result may be easily obtained. If $\partial\psi_1/\partial u$ and $\partial\psi_2/\partial u$ are nonzero at t^* and are of the same sign, then u is continuous at t^*.

To prove this, it is supposed first that $\partial\psi_1/\partial u > 0$ and $u(t^* + 0) = u(t^* - 0) - \epsilon, \epsilon > 0$. Taylor's theorem (Ref. 16, p. 56) implies

$$\psi_1 [t^*, u(t^* + 0)] = \psi_1 [t^*, u(t^* - 0)] - \epsilon \frac{\partial\psi_1 [t^*, u(t^* - 0) - \theta\epsilon]}{\partial u},$$

where $0 < \theta < 1$. But, $\partial\psi_1/\partial u > 0$ and $\epsilon > 0$, so

$$0 = \psi_1 [t^*, u + 0)] < \psi_1 [t^*, u(t^* - 0)] = 0$$

which contradicts the assumption $\epsilon > 0$. An identical argument holds in the remaining cases, so $u [t^*, u(t^* - 0)] = 0$.

7-4.3 STATICALLY INDETERMINATE PROBLEMS

A statically indeterminate beam may be classified as one of two types:

1. The beam is supported in such a way that all reactions are determined to within a finite number of unknown constants, or

2. The beam is supported in such a way that an infinite number of constants are required to specify the reactions (e.g., a beam on an elastic foundation).

In the first case, the unknown constants appear in the expressions for M and V. By defining new state variables, x_i with $i > 3$, to be these parameters, the following differential equations must be satisfied:

$$\frac{dx_i}{dt} = 0, i > 3.$$

In this way, statically indeterminate problems of the first type are reduced to variational problems to which Theorem 6-8 applies.

For statically indeterminate problems of the second type, however, more basic changes in formulation must be made. The fourth-order differential equation, Eq. 7-50 must be treated since $q(t)$ may be $q[t,x_1(t),x_2(t)]$. The fourth-order equation is equivalent to the first-order system

$$\frac{dx_1}{dt} = x_2$$

$$\frac{dx_2}{dt} = -\frac{x_3}{EI(u_i)}$$

$$\frac{dx_3}{dt} = x_4$$

and

$$\frac{dx_4}{dt} = -q(t,x_1,x_2)$$

where $x_1 = x$, $x_3 = M$, and $x_4 = V$. Theorem

6-8 now applies to statically indeterminate problems of the second type.

7-4.4 SOLUTION OF THE EQUATIONS OF THEOREM 7-1

The Lagrange multipliers, λ_0, λ_1, λ_2, are not uniquely determined by Theorem 7-1. However, if the solution, $x(t)$ and $u(t)$, is normal (Ref. 17, p. 214) for the problem, it is permissible to put $\lambda_0 = 1$. Theorem 7-1 then determines the remaining λ_i uniquely. Abnormal solutions are peculiar in that there may be no other functions, $x(t)$ and $u(t)$, near them which satisfy the conditions, Eqs. 7-50 through 7-55. The procedure adopted in this subparagraph is to assume there is a normal solution and then attempt to solve for it. If this fails, there is either no solution or the solution is abnormal, in which case a special analysis is required. In the following, λ_0 will be taken as 1.

In order to determine $u(t)$, consider any interval in which z $0 < z < 4$, of the inequalities, Eqs. 7-52 through 7-55, are equalities and the remaining $4 - z$ are strict inequalities. Eqs. 7-74 and 7-75 show that the $4 - z$ multipliers corresponding to the $4 - z$ strict inequalities are zero. Then, Eq. 7-73 is a system of m equations in the m functions $u_j(t)$ and the z nonzero multipliers. Further, the z equalities of Eqs. 7-52 through 7-55 yield z equations in the $u_j(t)$. Thus, there are $m + z$ equations which are to determine the $m + z$ unknowns. The nonzero $\mu_i(t)$ are first eliminated and the $u_j(t)$ are then found as functions of t and the parameters ξ_1 and ξ_2.

At points t_r^- one part of Eq. 7-55 is an equality, say the jth part ($j = 1$ or 2). In this case, only $\tau^-_{0,j+3,r}$ and $\tau^-_{1,j+3,r}$ can possibly be nonzero. Eqs. 7-77, 7-78, and 7-81 are three equations from which $\tau^-_{0,j+3,r}$ and $\tau^-_{1,j+3,r}$ can be eliminated and the t_r^- determined as functions of the parameters ξ_1 and ξ_2. Note that ξ_1 and ξ_2 to the left of t_r^- need not equal ξ_1 and ξ_2 to the right. In exactly the same way, the t_δ^- are determined by Eqs. 7-79, 7-80, and 7-82. Continuity of $\bar{H}$ at t_δ^+ and t_η^* determines the t_δ^+ and t_η^* as functions of ξ_1 and ξ_2.

The equations previously enumerated determine u_j, t_η^*, t_r^-, t_δ^-, and t_δ^+ as functions of t, ξ_1, and ξ_2. The problem would be solved by direct integration of Eqs. 7-71 and 7-72 and application of $g_1 = g_2 = 0$ if ξ_1 and ξ_2 were known.

The conditions which determine ξ_1 and ξ_2 are Eqs. 7-83, 7-84, 7-85, and 7-86. Eqs. 7-85 and 7-86 after elimination of γ_1 and γ_2, yield two equations in ξ_1 and ξ_2. If there are w of the points t_r^- and t_δ^-, they subdivide $(0,T)$ into $w + 1$ subintervals. Since these are the only possible points of discontinuity of λ_1 and λ_2, there are just $w + 1$ pairs, ξ_1 and ξ_2, which represent $2w + 2$ unknown. Thus, there are $2w + 2$ equations from which the $2w + 2$ values of ξ_1 and ξ_2 may be determined. If this is not the case, the problem has no solution or the solution is abnormal.

It is assumed now that $u(t)$ and points t_r^- and t_δ^- are known functions of ξ_1 and ξ_2. A numerical method is developed which can be used to solve the equations given above for ξ_1 and ξ_2. A numerical solution is required since, even for very simple problems, the function $f_2(t,\xi_1,\xi_2)$ is far too complicated to integrate in closed form.

Expressions, Eqs. 7-85 and 7-86 generally yield two easy relations between ξ_1 and ξ_2. Eqs. 7-83 and 7-84, however, require successive integration of $f_2(t,\xi_1,\xi_2)$. To complicate matters, some limits of integration are the points t_r^- and t_s^-, which are themselves functions of ξ_1 and ξ_2. Therefore, Eqs. 7-83 and 7-84 form a system of nonlinear, finite (nondifferential) equations in ξ_1 and ξ_2.

A Generalized Newton Method (Ref. 18, p. 220) is used to solve this set of equations. A generalization of Leibniz' Rule is used to compute the required derivatives of integrals with variable limits of integration. This rule is (Ref. 16, p. 80).

$$\left.\begin{aligned} \frac{d}{d\tau}\int_{g_1(\tau)}^{g_2(\tau)} f_2(t,\tau)dt \\ = \int_{g_1(\tau)}^{g_2(\tau)} \frac{\partial f_2(t,\tau)}{\partial \tau}\,dt \\ + f_2\,[g_2(\tau),\tau]\,\frac{dg_2(\tau)}{d\tau} \\ - f_2\,[g_1(\tau),\tau]\,\frac{dg_1(\tau)}{d\tau} \end{aligned}\right\} \quad (7\text{-}87)$$

where $\partial f_2(t,\tau)/\partial\tau$ is piecewise continuous, and g_1 and g_2 are differentiable. It is assumed, also, that f_2 and $\partial f_2/\partial\tau$ are continuous at $t = g_1(\tau)$ and $g_2(\tau)$.

7-4.5 BEAMS WITH RECTANGULAR CROSS SECTION OF VARIABLE DEPTH

Three examples are considered in this subparagraph. In each example, the beam cross section is rectangular with fixed width and variable depth. Also, the constraint, Eq. 7-55, is taken as

$$-\Delta < x(t) < \Delta \qquad (7\text{-}88)$$

where $\Delta > 0$ is a constant.

First, the equations of Theorem 7-1 are written out in detail and simplified for the case of beam with rectangular cross section of variable depth. In pars. 7-4.5.1, 7-4.5.2, and 7-4.5.3, three specific examples are considered. These examples range from an easy problem in par. 7-4.5.1 to a rather complex one in par. 7-4.5.3.

The cross section considered here is shown in Fig. 7-9. For this particular cross section, if

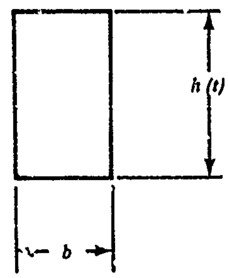


Figure 7-9. Rectangular Cross Section

$\tau_{max} > 1/2\ \sigma_{max}$, Eqs. 7-52 and 7-53 are satisfied if and only if Eqs. 7-56 and 7-57 are satisfied. This result may be proved by expressing σ_p and τ_p as functions of d and applying methods of ordinary calculus. The restriction $\tau_{max} > 1/(2\ \sigma_{max})$ is necessary, since at the extreme fiber of the beam, the principal shear stress is half the principal normal stress. This relation between τ_{max} and σ_{max} is no restriction for design of metallic beams. Yield stresses for steel and other common metals satisfy this condition.

In this case, there is only one design variable $h(t)$; Eqs. 7-44 through 7-48 are

$$I(t) = \frac{h}{12}h^3(t) \qquad (7\text{-}89)$$

$$A(t) = bh(t) \qquad (7\text{-}90)$$

$$b(t) = b \tag{7-91}$$

$$Q(t) = \frac{b}{8} h^2(t) \tag{7-92}$$

and

$$C(t) = \frac{1}{2} h(t). \tag{7-93}$$

$M(t)$ and $V(t)$ are assumed to be known, piecewise twice continuously differentiable functions of t whose discontinuities occur at points $t = \omega_\alpha$.

Eqs. 7-89 through 7-93 and Eq. 7-70 along with $\lambda_0 = 1$, yield

$$\bar{H} = -bh + \lambda_1 x_2 - \lambda_2 \left[\frac{12M(t)}{Eb}\right] h^{-3}. \tag{7-94}$$

The procedure outlined is now used to determine $h(t)$. In any interval where Eq. 7-55 is an equality,

$$\frac{6|M(t)|}{bh^2(t)} = \sigma_{max}$$

so

$$h(t) = \left[\frac{6|M(t)|}{b\sigma_{max}}\right]^{1/2}.$$

In any interval where Eq. 7-57 is an equality,

$$\frac{3|V(t)|}{2bh(t)} = \tau_{max}$$

so

$$h(t) = \frac{3|V(t)|}{2b\tau_{max}}.$$

In any interval where Eq. 7-54 is an equality,

$$\frac{bh^3(t)}{12} = I_0$$

so

$$h(t) = \left(\frac{12I_0}{b}\right)^{1/3}. \tag{7-95}$$

In any interval, (t_δ^-, t_δ^+), where Eq. 7-88 is an equality, direct differentiation and use of Eqs. 7-71 and 7-72 yield

$$0 = -\frac{12M(t)}{Eh^2(t)}. \tag{7-96}$$

In order for Eq. 7-96 to be satisfied, it is necessary that $M(t) = 0$ (hence also $V(t) = 0 = q(t)$) must be identically satisfied in (t_δ^-, t_δ^+). If this is the case, $h(t)$ is given by Eq. 7-95.

In any interval where Eqs. 7-56, 7-57, 7-54, and 7-88 are all strict inequalities, Eqs. 7-84 and 7-75 show that $\mu_i(t) = 0, i = 1, ..., 5$. Eq. 7-73 then is

$$-b + \lambda_2(t)\left[\frac{36M(t)}{Eb}\right] h^{-4}(t) = 0 \tag{7-97}$$

so

$$h(t) = \left[\frac{36\lambda_2(t)M(t)}{Eb^2}\right]^{1/4}.$$

It is worthwhile to note that in order for Eq. 7-97 to hold the product $\lambda_2(t)M(t)$ must be positive. That is, $\lambda_2(t)$ and $M(t)$ must have the same algebraic sign throughout any interval in which Eq. 7-97 holds.

In a more compact notation,

$$h(t) = \begin{cases} \dfrac{3|V(t)|}{2b\tau_{max}}, & \text{if } |\tau| = \tau_{max} \\ \left(\dfrac{6|M(t)|}{b\sigma_{max}}\right)^{1/2}, & \text{if } |\sigma| = \sigma_{max} \\ \left(\dfrac{12I_0}{b}\right)^{1/3}, & \text{if } I = I_0 \\ \left[\dfrac{36\lambda_2(t)M(t)}{Eb^2}\right]^{1/4}, & \text{if } |\tau| < \tau_{max},\ |\sigma| < \sigma_{max},\ \text{and } I > I_0. \end{cases} \tag{7-98}$$

The Weierstrass condition shows that the largest of the expressions in Eq. 7-98 is the proper value of $h(t)$.

Eq. 7-98 in Eq. 7-64 yields

$$f(t) = \begin{cases} -C_1 M(t)|V(t)|^{-3}, & \text{if } |\tau| = \tau_{max} \\ -C_2|M(t)|^{-1/2} \times \operatorname{sgn}[M(t)], & \text{if } |\sigma| = \sigma_{max} \\ -\dfrac{M(t)}{EI_0}, & \text{if } I = I_0 \\ -C_3|\lambda_2(t)|^{-3/4} \times |M(t)|^{1/4} \times \operatorname{sgn}[M(t)], & \text{if } |\tau| < \tau_{max},\ |\sigma| < \sigma_{max},\ \text{and } I > I_0, \end{cases} \tag{7-99}$$

where

$$C_1 = \frac{32b^2\tau^3_{max}}{9E}$$

$$C_2 = \left(\frac{2b\sigma^3_{max}}{3E^2}\right)^{1/2}$$

$$C_3 = \left(\frac{4b^2}{9E}\right)^{1/4}$$

and the function sgn ( ) is defined by the relation

$$q\ \operatorname{sgn}(q) = q$$

for real q.

Equations which determine the special points t_r^-, t_δ^+, t_δ^-, and t_η^* may now be found. In the problem at hand, $X_1(t)$ and $X_2(t)$ are constant, so their derivatives are zero. Eq. 7-81 is then

$$\bar{H}(t_r^- - 0) = \bar{H}(t_r^- + 0)\,. \tag{7-100}$$

Experience has shown that on both sides of t_r^-, Eqs. 7-56, 7-57, and 7-54 are strict inequalities. Assuming this is this the case, Eqs. 7-94 and 7-98 together with Eq. 7-83 or 7-84 may be used to simplify Eq. 7-100. The result is

$$\left[\lambda_2(t_r^- - 0)M(t_r^- - 0)\right]^{1/4} = \left[\lambda_2(t_r^- + 0)M(t_r^- + 0)\right]^{1/4}\,. \tag{7-101}$$

Eq. 7-98 and $I > 0$ imply $M(t_r^-) \neq 0$, so if $M(t)$ is continuous at t_r^-, then Eq. 7-101 reduces to

$$\lambda_2(t_r^- - 0) = \lambda_2(t_r^- + 0)\,. \tag{7-102}$$

Points of discontinuity of $M(t)$ must be checked in Eq. 7-101 as possible t_r^-.

Eq. 7-96 shows that points t_δ^+ and t_δ^- can occur only in intervals where $M(t)$ (hence also

$V(t)$ and $q(t)$) is identically zero. Since this situation is not common in practical problems, such intervals will not be discussed here.

According to Theorem 7-1, the points t_η^* are determined by the condition

$$\bar{H}(t_\eta^* - 0) = \bar{H}(t_\eta^* + 0) . \tag{7-103}$$

By direct computation it is seen that the partial derivatives with respect to h of the left sides of

$$|\tau| - \tau_{max} < 0$$

$$|\sigma| - \sigma_{max} < 0$$

and

$$I_l - I < 0$$

are all negative at points where $M(t) \neq 0 \neq V(t)$. The result stated just below Corollary 7-1 then shows that points of intersection of intervals in which one or more of the above inequalities is an equality may be determined by the condition

$$h(t_\eta^* - 0) = h(t_\eta^* + 0) . \tag{7-104}$$

If Eq. 7-104 is used to determine t_η^*, then points ω_α must be checked in Eq. 7-103 as possible t_η^*.

Let $Q = t_\eta^* \neq \omega_\alpha$ be defined to be a point of intersection of two intervals such that $|\tau| = \tau_{max}$ on one side of $t = Q$ and Eqs. 7-56, 7-57, 7-54, and 7-88 are strict inequalities on the other. Point Q is to be determined by Eq. 7-103. Due to continuity at $t = Q$, Eq. 7-103 may be written as

$$- b\left(\frac{3}{2b\tau_{max}}\right)|V(Q)|$$

$$- C_1\lambda_2(Q)M(Q)\,|V(Q)|^{-3}$$

$$= - b\left(\frac{36}{Eb^2}\right)^{1/4} [\lambda_2(Q)M(Q)]^{1/4}$$

$$- C_3[\lambda_2(Q)M(Q)]^{1/4}$$

Using the definition of C_3 in this equation and manipulating the result yields

$$\frac{3|V(Q)|^4}{\lambda_2(Q)M(Q)} - 4\left(\frac{64b^2\tau_{max}^4}{9E}\right)^{1/4}$$

$$\times\left[\frac{|V(Q)|^4}{\lambda_2(Q)M(Q)}\right]^{3/4} + \frac{64b^2\tau_{max}^4}{9E} = 0. \tag{7-105}$$

By putting

$$P = \left[\frac{|V(Q)|^4}{\lambda_2(Q)M(Q)}\right]^{1/4}$$

and

$$C = \left(\frac{64b^2\tau_{max}^4}{9E}\right)^{1/4}$$

Eq. 7-105 becomes

$$3P^4 - 4CP^3 + C^4 = 0.$$

The roots of this equation in P are C, C, $C(1 + \sqrt{2}i)$, and $C(1 - \sqrt{2}i)$, where $i^2 = -1$. The fourth powers of the last two roots are not real, so the only real solution of Eq. 7-103 is

$$|V(Q)|^4 = \left(\frac{64b^2\tau_{max}^4}{9E}\right)\lambda_2(Q)M(Q) . \tag{7-106}$$

Similarly, $S = t^*_\eta \neq \omega_\alpha$ is defined to be a point of intersection of two intervals such that $|\sigma| = \sigma_{max}$ on one side of S and Eqs. 7-56, 7-57, 7-54, and 7-88 are strict inequalities on the other. Just as above, Eq. 7-103 reduces to

$$3\hat{P} - 4\hat{C}\hat{P} + \hat{C}^4 = 0$$

where

$$\hat{P} = \left[\frac{M(S)}{\lambda_2(S)}\right]^{1/4}$$

and

$$\hat{C} = \left(\frac{\sigma^2_{max}}{E}\right)^{1/4}$$

Therefore, the only real solution of Eq. 7-103 at $t = S$ is

$$M(S) = \left(\frac{\sigma^2_{max}}{E}\right)\lambda_2(S). \qquad (7\text{-}107)$$

In deriving Eqs. 7-106 and 7-107, it was assumed that Q and S were not equal to any ω_α. The ω_α are, therefore, possible choices for Q and S and must be checked in Eq. 7-103.

In particular problems, the following two identities are used:

$$\int_A^C \int_A^\nu f(\eta)d\nu = (C - B)\int_A^B f(\eta)d\eta + \int_A^B \int_A^\nu f(\eta)d\eta d\nu + \int_B^C \int_B^\nu f(\eta)d\eta d\nu, \qquad (7\text{-}108)$$

where $A < B < C$, and

$$\frac{d}{d\alpha}\left[\int_{g_1(\alpha)}^{g_2(\alpha)} \int_{g_1(\alpha)}^{\nu} f(\eta,\alpha)d\eta d\nu\right] = \int_{g_1(\alpha)}^{g_2(\alpha)} \int_{g_1(\alpha)}^{\nu} \frac{\partial f(\eta,\alpha)}{\partial \alpha} d\eta d\nu - [g_2(\alpha) - g_1(\alpha)] \times f[g_1(\alpha), \alpha]\frac{dg_1(\alpha)}{d\alpha} + \left[\int_{g_1(\alpha)}^{g_2(\alpha)} f(\eta,\alpha)d\eta\right]\frac{dg_2(\alpha)}{d\alpha}. \qquad (7\text{-}109)$$

Leibniz' Rule, Eq. 7-87, is used repeatedly to obtain Eq. 7-109.

7-4.5.1 A PROBLEM WHICH CAN BE SOLVED ANALYTICALLY

As a first example, the cantilever beam of Fig. 7-10 is considered. This problem is simple enough that a solution can be obtained analytically.

Boundary conditions for this beam are

$$x_1(0) = x_2(0) = 0. \qquad (7\text{-}110)$$

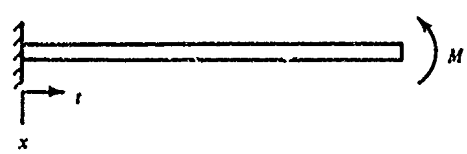


Figure 7-10. Simple Cantilever Beam

The bending moment and shear are

$$M(t) = M > 0 \quad (M \text{ constant})$$

and

$$V(t) = 0.$$

For simplicity, let $I_0 > 0$ be small enough so that

$$\left(\frac{12I_0}{b}\right)^{1/3} < \left(\frac{6M}{b\sigma_{max}}\right)^{1/2}.$$

If this is the case then $I(t) > I_0$ for all t. Further, since $\tau(t) = 0$, Eq. 7-98 may be simplified to

$$h(t) = \begin{cases} \left[\dfrac{6M}{b\sigma_{max}}\right]^{1/2}, & \text{if } |\sigma| = \sigma_{max} \\ \left[\dfrac{36\lambda_2(t)M}{Eb^2}\right]^{1/4}, & \text{if } |\sigma| < \sigma_{max}. \end{cases} \tag{7-111}$$

Also, Eq. 7-99 becomes

$$f_2(t) = \begin{cases} -\left(\dfrac{2b\sigma^3_{max}}{3E^2M}\right)^{1/2}, & \text{if } |\sigma| = \sigma_{max} \\ -\left(\dfrac{4b^2M}{9E}\right)^{1/4} |\lambda_2(t)|^{-3/4}, & \text{if } |\sigma| < \sigma_{max}. \end{cases} \tag{7-112}$$

Integration of the differential equations

$$\frac{dx_1}{dt} = x_2$$

and

$$\frac{dx_2}{dt} = f_2(t)$$

and application of Eq. 7-110 yields

$$x_2(t) = \int_0^t f_2(\eta)d\eta \tag{7-113}$$

and

$$x_1(t) = \int_0^t \int_0^v f_2(\eta)d\eta dv. \tag{7-114}$$

The inequality $f_2(t) < 0$ and Eq. 7-113 imply $x_2 < 0$ in $(0,T)$, so there can be no points t_r^-. The only possible point at which Eq. 7-88 is an equality is $t = T$. Since $f_2(t) < 0$, Eq. 7-114 implies $x_1(t) < 0$. Therefore, if Eq. 7-88 is an equality, it must be $x_1(T) = -\Delta$. Further, since there are no t_r^-, t_δ^+, t_c^-, or ω_α, λ_1 and λ_2 are continuous and there is only one pair of constants, ξ_1 and ξ_2, to be determined.

It is assumed first that $x_1(T) > -\Delta$. In this case, Eq. 7-86 implies

$$\lambda_1(T) = \xi_1 = 0$$

and

$$\lambda_2(T) = \xi_2 - T\xi_1 = 0$$

which in turn implies $\xi_1 = \xi_2 = 0$. Since $\lambda_2 = 0$ throughout $(0,T)$, the second part of Eq. 7-111 cannot occur. Therefore, the beam of minimum weight is uniform. Eqs. 7-112 and 7-114 then yield

$$x_1(T) = -\left(\frac{2b\sigma^3_{max}}{3E^2M}\right)^{1/2} \frac{T^2}{2}$$

Therefore, Eq. 7-111 with $|\sigma| = \sigma_{max}$ is the solution of the problem provided the deflection requirement Δ is such that

$$\Delta > \left(\frac{b\sigma^3_{max}T^4}{6E^2M}\right)^{1/2} . \tag{7-115}$$

If the deflection requirement Δ does not satisfy Eq. 7-115, then it is necessary that $x_1(T) = -\Delta$. Otherwise, this argument would hold, and the deflection at T would violate Eq. 7-88. Therefore, the additional boundary condition,

$$g_3 = x_1(T) + \Delta = 0,$$

must be satisfied. The two constants ξ_1 and ξ_2 must now be found.

The only useful relation given by Eq. 7-86 is

$$\lambda_2(T) = \xi_2 - \xi_1 T = 0.$$

This implies $\xi_2 = \xi_1 T$, so that

$$\lambda_2(T) = \xi_1(T - t),$$

and only ξ_1 remains to be found.

On physical grounds, it is expected that the beam should be stiffest near $t = 0$ in order to reduce the deflection at T efficiently. Also, since λ_2 is largest at $t = 0$, the second part of Eq. 7-111 would tend to stiffen the beam there. It is assumed, therefore, that there is just one point t^* having $|\sigma| < \sigma_{max}$ on its left and $|\sigma| = \sigma_{max}$ on its right. Eq. 7-107 for t^* is

$$M = \left(\frac{\sigma^2_{max}}{E}\right)\xi_1(T - t^*) . \tag{7-116}$$

Eq. 7-114 with $t = T$ may be integrated using Eqs. 7-108 and 7-112 and becomes

$$\begin{aligned} x_1(T) = {} & (T - t^*)4\left(\frac{4b^2M}{9E}\right)^{1/4} \\ & \times \xi_1^{-3/4}\left[(T - t^*)^{1/4} - T^{1/4}\right] \\ & - 4\left(\frac{4b^2M}{9E}\right)^{1/4} \\ & \times \xi_1^{-3/4}\left[\frac{4}{5}(T - t^*)^{5/4} + T^{1/4}\, t^*\right] \\ & - \left(\frac{2b\sigma^3_{max}}{3E^2M}\right)^{1/2} (T - t^*)^2 . \end{aligned}$$

The right-hand side of this equation may be simplified by eliminating either ξ_1 or t^* through use of Eq. 7-116. Since ξ_1 does not have as much physical significance as t^*, it is eliminated. The conditions $x_1(T) = -\Delta$ becomes

$$-\Delta = \left(\frac{2b\sigma^3_{max}}{3E^2M}\right)^{1/2} \times \left[\frac{3}{10}(T - t^*)^2 - \frac{4}{5}T^{5/4}(T - t^*)^{3/4}\right] \tag{7-117}$$

The derivative of the right side of Eq. 7-117 with respect to t^* is zero at $t^* = T$ and positive everywhere else. This means that Eq. 7-117 has at most one solution. Eq. 7-116 then determines ξ_1 and the problem is solved.

As a numerical example, the beam of Fig. 7-10, having the following properties, is considered:

$T = 10$ in.

$b = 1$ in.

$\sigma_{max} = 30{,}000$ lb/in.^{2}

$E = 10^7$ lb/in.^{2}

and

$M = 450$ in.-lb

If $\Delta > 1$ in., then the beam of minimum weight is uniform with $h = 0.30$ in.

For a more meaningful problem, $\Delta = 0.5$ in. is considered. Eqs. 7-117 and 7-116 yield $t^* = 7.7$ in. and $\xi_1 = 2.17$. The precise shape of the optimal beam is given by Eq. 7-111. By putting Eqs. 7-111 and 7-90 in Eq. 7-49 and performing the indicated integration, the volume of the optimal beam is found to be 3.59 in.^{3} A plot of the profile of the optimal beam may be made by direct substitution into Eq. 7-111. This profile is shown in Fig. 7-11.

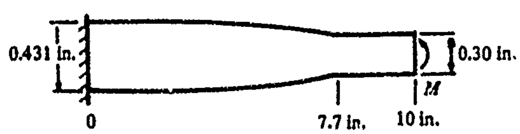


Figure 7-11. Cantilever Beam of Minimum Weight

By elementary computation, it is seen that the uniform beam which has $x_1(10) = -0.5$ in. and satisfies Eqs. 7-56 and 7-57 is 0.378 in. deep; so its volume is 3.78 in.^{3} The designed beam, therefore, has 5.3% less volume than a uniform beam which will satisfy the same stress and deflection requirements.

7-4.5.2 SIMPLY SUPPORTED BEAM WITH POSITIVE DISTRIBUTED LOAD

The beam considered here is simply supported (see Fig. 7-12) with piecewise continuously differentiable distributed load $q(t)$ $q(t) > 0$ for all t in $(0,T)$. The load $q(t)$ of this form implies $V(t)$ is non-negative at zero and decreases monotonically in $(0,T)$. $M(t)$ is zero at both $t = 0$ and $t = T$. Further, on either side of the point where $M(t)$ has its (non-negative) maximum it is monotone.

Figure 7-12. Simply Supported Beam With Positive Distributed Load

If $\int_0^T q(t)dt \neq 0$, then $M(t)$ and $V(t)$ cannot be zero at the same point. There is, therefore, no danger that the optimal beam will have $h(t) = 0$ at any point. For this reason, the constraint, Eq. 7-54, is not imposed here.

Since $M(0) = M(T) = 0$ and $V(0) \neq 0 \neq V(t)$, $|\tau| = \tau_{max}$ is expected near the ends of the beam. Toward the center of the beam, $M(t)$ becomes large, so $|\sigma| = \sigma_{max}$ is expected there. If the beam requires stiffening, the additional material can best be used near the point of maximum deflection. This argument indicates that $(0,T)$ should be broken up into subintervals as shown in Fig. 7-13.

In terms of previous notation, t_1, t_2, t_4, and t_5 correspond to the notation t_η^*; and t_3 to t_r^*. For certain ranges of Δ some of the subintervals shown in Fig. 7-13 will not appear.

Provided t_1 and t_5 separate intervals in which $|\tau| = \tau_{max}$ and $|\sigma| = \sigma_{max}$, they are determined by Eq. 7-104 The points t_2 and t_4, when they exist, are determined by Eq. 7-107. Finally, t_3 is determined by Eq. 7-102.

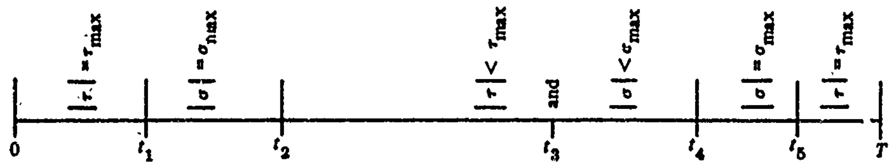


Figure 7-13. Subdivision of the Beam With Distributed Load

The boundary conditions for this problem are

$$x_1(0) = 0$$

and

$$x_1(T) = 0.$$

Eqs. 7-75 and 7-76, therefore, yield

$$\lambda_2(0) = \lambda_2(T) = 0.$$

The constants ξ_1 and ξ_2 in λ_2 may have different values on opposite sides of t_3. To the left of t_3,

$$\lambda_2(0) = \xi_2 - \xi_1 \times 0 = 0$$

so

$$\lambda_2(t) = -\xi_1 t = \zeta_1 t \tag{7-118}$$

To the right of t_3,

$$\lambda_2(T) = \xi_2 - \xi_1 T = 0$$

so

$$\lambda_2(t) = \xi_2\left(1 - \frac{t}{T}\right) = \zeta_2\left(1 - \frac{t}{T}\right). \tag{7-119}$$

The constants ζ_1 and ζ_2, which are introduced in Eqs. 7-118 and 7-119, are now to be determined.

Eq. 7-102 for t_3 yields

$$t_3 = \frac{T\zeta_2}{T\zeta_1 + \zeta_2}.$$

Assuming t_2 and t_4 exist, the equations that determine them are

$$M(t_2) = \left(\frac{\sigma^2_{max}}{E}\right) \zeta_1 t_2 \tag{7-120}$$

and

$$M(t_4) = \left(\frac{\sigma^2_{max}}{E}\right) \zeta_2 \left(1 - \frac{t^4}{T}\right) \tag{7-121}$$

In this case, t_1 and t_5 are determined by Eq. 7-105. If t_2 or t_4 does not exist, then t_1 or t_5 is determined by

$$|V(t_1)|^4 = \left(\frac{64 b^2 \tau^4_{max}}{9E}\right) \zeta_1 t_1 M(t_1) \tag{7-122}$$

or

$$|V(t_5)|^4 = \left(\frac{64 b^2 \tau^4_{max}}{9E}\right) \times \zeta_2 \left(1 - \frac{t_5}{T}\right) M(t_5). \tag{7-123}$$

It is noted that Eqs. 7-120 through 7-123 can be solved easily for ζ_1 and ζ_2, but, in general, not so easily for the t_i. In the development which follows, it will be convenient to use Eqs. 7-120 through 7-123 to solve for ζ_1 and ζ_2 as functions of the t_i.

The conditions that are to determine the unknown t_i are $x_1(t_3) = \Delta$ and $x_2(t_3) = 0$. However, for computational reasons, a more convenient set of equivalent conditions is

$$x_1(0) = 0 \tag{7-124}$$

and

$$x_1(T) = 0 \tag{7-125}$$

where $x_1(t_3) = \Delta$ and $x_2(t_3) = 0$ are used as initial conditions for integration.

Conditions, Eqs. 7-124 and 7-125 may be written explicitly as

$$R_1 = \Delta + t_1 \int_0^{t_1} f_2(\eta)d\eta + t_2 \int_{t_1}^{t_2} f_2(\eta)d\eta$$

$$+ t_3 \int_{t_2}^{t_3} f_2(\eta,\zeta_1)d\eta$$

$$- \int_0^{t_1} \int_0^{\nu} f_2(\eta)d\eta d\nu$$

$$- \int_{t_1}^{t_2} \int_{t_1}^{\nu} f_2(\eta)d\eta d\nu$$

$$- \int_{t_2}^{t_3} \int_{t_2}^{\nu} f_2(\eta,\zeta_1)d\eta d\zeta = 0 \tag{7-126}$$

and

$$R_2 = \Delta + (T - t_4) \int_{t_3}^{t_4} f_2(\eta,\zeta_2)d\eta$$

$$+ (T - t_5) \int_{t_4}^{t_5} f_2(\eta)d\eta$$

$$+ \int_{t_3}^{t} \int_{t_3}^{\nu} f_2(\eta,\zeta_2)d\eta d\nu$$

$$+ \int_{t_4}^{t_5} \int_{t_4}^{\nu} f_2(\eta)d\eta d\nu$$

$$+ \int_{t_5}^{T} \int_{t_5}^{\nu} f_2(\eta)d\eta d\nu = 0 , \tag{7-127}$$

where P_1 and R_2 are introduced for notational purposes.

It is assumed now that $q(t)$, b, E, T, σ_{max}, and τ_{max} are given. The equations that determine the t_i are different in four distinct ranges of the deflection requirement. These ranges of Δ are described in the following:

1. Δ_0 denotes the largest deflection that occurs when the subinterval (t_2, t_4) does not appear, i.e., when the beam is specified by only the first two parts of Eq. 7-98.

If $\Delta > \Delta_0$, then the beam specified by the first two parts of Eq. 7-98 is the one of minimum weight.

2. For Δ slightly less than Δ_0, there exist points t_2 and t_4 that are determined by Eqs. 7-126 and 7-127.

As Δ decreases, points t_2 and t_4 move toward zero and T, respectively. There is a value of Δ, say $\Delta = \Delta_1$, for which either t_2 or t_4 first coincides with t_1 or t_5, respectively. For definiteness, assume $t_2 = t_1$ when $\Delta = \Delta_1$.

3. For Δ slightly less than Δ_1, points t_1 and t_4 are determined by Eqs. 7-126 and 7-127.

As Δ decreases, points t_1 and t_4 move toward 0 and T, respectively. There is a value of Δ, say $\Delta = \Delta_2$, for which t_4 first coincides with t_5.

4. For $\Delta < \Delta_2$, points t_1 and t_5 are determined by Eqs. 7-126 and 7-127.

This explanation of the behavior of the t_i is not the result of a mathematical analysis. It is expected on physical grounds and has been valid in each case treated.

The values of Δ_1 and Δ_2 could be obtained analytically. However, their determination would be of the same order of difficulty as the optimization problem considered in this paragraph.

Δ_1 and Δ_2 can be determined by a trial and error scheme. For example, to determine Δ_1, t_2 is put equal to t_1 and Eq. 7-107 determines ζ_1. Then, t_3 is guessed and Eq. 7-102 solved for ζ_2. Numerical integration of Eqs. 7-60 and 7-61 indicates the correction that is to be made in t_3. When t_3 is located accurately, the resulting deflecting at t_3 is Δ_1. A similar procedure is used to determine Δ_2 (t_4 is put equal to t_5).

The solution of Eqs. 7-126 and 7-127 for the t_i differs only in certain details depending on whether the given value of Δ is in the range described by Cases 2, 3, or 4. The method of solution will be described for the Case 2.

An iterative method called Generalized Newton Method (Ref. 18) is used to solve for t_2 and t_4. The procedure begins by estimating values $\hat{t}_2$ and $\hat{t}_4$; and then making a correction according to the formula

$$\begin{bmatrix} \bar{t}_2 \\ \\ \bar{t}_4 \end{bmatrix} = \begin{bmatrix} \hat{t}_2 \\ \\ \hat{t}_4 \end{bmatrix} - \begin{bmatrix} \dfrac{\partial R_1}{\partial t_2} & \dfrac{\partial R_1}{\partial t_4} \\ \\ \dfrac{\partial R_2}{\partial t_2} & \dfrac{\partial R_2}{\partial t_4} \end{bmatrix}^{-1} \begin{bmatrix} R_1 \\ \\ R_2 \end{bmatrix} \tag{7-128}$$

where $[\;]^{-1}$ denotes matrix inverse, and $\bar{t}_2$ and $\bar{t}_4$ are improvements on the estimate.

Eqs. 7-102 and 7-107 determine $t_3 = t_3(\zeta_1, \zeta_2)$, $\zeta_1 = \zeta_1(t_2)$, and $\zeta_2 = \zeta_2(t_4)$. By use of this information, the derivatives in Eq. 7-128 are computed by the chain rule of differentiation. For example,

$$\frac{\partial R_1}{\partial t_2} = R_{1,t_2} + \left(\frac{\partial R_1}{\partial \zeta_1} + \frac{\partial R_1}{\partial t_3} \frac{\partial t_3}{\partial \zeta_1} \right) \frac{\partial d\zeta_1}{dt_2} \tag{7-129}$$

where

$$R_{1,t_2} = t_3 [f_2(t_2 - 0) - f_2(t_2 + 0)]$$

is the partial derivative of R_1 with respect to t_2 with all variables in R_1 taken as independent,

$$\frac{\partial R_1}{\partial \zeta_1} = t_3 \int_{t_1}^{t_2} \frac{\partial f_2(\eta, \zeta_1)}{\partial \zeta_1} d\eta$$

$$- \int_{t_3}^{t_3} \int_{t_3}^{v} \frac{\partial f_2(\eta,\zeta_1)}{\partial \zeta_1} d\eta dv$$

and

$$\frac{\partial R_1}{\partial t_3} = t_3 f_2 (t_3 - 0).$$

Similar expressions for the remaining derivatives in Eq. 7-128 are derived with the aid of Eq. 7-109.

The iterative procedure for determining t_2 and t_4 is:

Step 1. Make an estimate, $\hat{t}_2$ and $\hat{t}_4$,

Step 2. Solve Eqs. 7-120 and 7-121 for ζ_1 and ζ_2,

Step 3. Compute, numerically, all the integrals in Eqs. 7-127, 7-128, and the remaining derivatives corresponding to Eq. 7-129,

Step 4. Compute the right side of Eq. 7-128, and

Step 5. With this improved estimate return to Step 1.

This procedure has been programmed for a digital computer. The program was arranged in such a way that only Δ_0, Δ_1, Δ_2, $M(t)$, $V(t)$, and the physical properties of the beam need to be specified. Many different loading situations may thus be considered without altering the program appreciably.

As a numerical example, a beam with the following properties is considered:

$q(t)$ = t lb/in.

T = 40 in.

b = 0.25 in.

E = 10^7 lb/in.^{2}

τ_{max} = 15,000 lb/in.^{2}

and

σ_{max} = 30,000 lb/in.^{2}

For this problem, it was found that

Δ_0 = 0.774 in.

Δ_1 = 0.728 in. ($t_2 = t_1$)

and

Δ_2 = 0.470 in. ($t_4 = t_5$).

As has been noted, the magnitude of the deflection requirement Δ plays a major role in the outcome of a particular problem. In order to emphasize the effect of Δ on the properties of the optimal beam, the numerical example given was solved for eight different values of Δ. The results are presented in Table 7-4.

In Table 7-4, the first column contains the values of Δ considered. The following seven columns give information which, when substituted into Eq. 7-98, completely specifies the optimal beam. The next column gives the volume of this optimal beam. The final two columns give the volume of the lightest beam of constant depth which satisfies the conditions of the problem and the percent saving realized when the optimal beam is used instead of this uniform beam. Dashes have been inserted in the table when the quantity to be tabulated does not exist.

TABLE 7-4
RESULTS FOR SIMPLY SUPPORTED BEAM WITH $q(t) = t$

Δ, in.	t_1, in.	t_2, in.	t_3, in.	t_4, in.	t_5, in.	ξ_1	ξ_2	Vol., in.^{3}	Vol. of Unif. Beam[1], in.^{3}	Saving, %
>0.775	0.053	----------	----------	----------	39.89	--------	--------	14.00	18.15	29.6
0.75	0.053	10.21	10.42	22.54	39.89	2.77	104.7	14.1[illegible]	18.15	28.6
0.65	0.048	0.048	19.99	28.41	39.89	3.60	144.0	14.75	18.15	23.0
0.50	0.040	0.040	20.32	37.79	39.89	5.27	217.7	16.03	18.58	15.5
0.40	0.034	0.034	20.32	39.90	39.90	7.12	294.1	17.[illegible]3	20.01	15.4
0.30	0.028	0.028	20.32	39.92	39.92	10.45	431.7	19.[illegible]7	22.03	15.5
0.20	0.022	0.022	20.32	39.94	39.94	17.91	740.0	31.82	25.21	15.5
0.10	0.014	0.014	20.32	39.96	39.96	45.17	1866.0	27.50	31.78	15.6

[1] Volume of lightest beam of constant depth which satisfies all the requirements of the problem.

For each value of Δ, the iterative procedure used to solve the problem required approximately 40 sec per iteration on an IBM 1410 Computer. Further, three to six iterations were sufficient to obtain convergence of the residue to seven decimal places, so computing time was not excessive.

It is shown (Ref. 18, p. 222) that if the sequence of approximations constructed by the Generalized Newton Algorithm converges, then it must converge quadratically, i.e., the error at the $n + 1$st step is proportional to the square of the error at the nth step. This rapid convergence was observed in the numerical calculations and explains why only three to six iterations were required.

7-4.5.3 A PROBLEM OF A MORE GENERAL TYPE

The beam considered here is loaded as shown in Fig. 7-14.

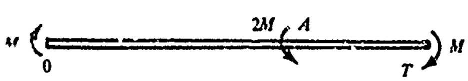


Figure 7-14. Beam With an Inflection Point

Boundary conditions are, as in the simply supported case,

$$x_1(0) = x_1(T) = 0.$$

For the given loading,

$$M(t) = \begin{cases} M, & \text{if } 0 < t < A \\ -M, & \text{if } A < t < T \end{cases}$$

and

$$V(t) = 0.$$

In this problem $A > T/2$ is assumed.

Since $M(t)$ is never zero, the requirement $|\sigma| < \sigma_{max}$ implies $h \neq 0$ for all t. $I = 0$ is impossible, so the requirement $I > I_0$ is not enforced.

Eqs. 7-98 and 7-99 in this case are

$$h(t) = \begin{cases} \left(\dfrac{6M}{b\sigma_{max}}\right)^{1/2}, & \text{if } |\sigma| = \sigma_{max} \\ \left[\dfrac{36\lambda_2(t)M(t)}{Eb^2}\right]^{1/4}, & \text{if } |\sigma| < \sigma_{max} \end{cases} \qquad (7\text{-}150)$$

and

$$f_2(t) = \begin{cases} -C_2 M^{-1/2} \times \operatorname{sgn}[M(t)], & \text{if } |\sigma| = \sigma_{max} \\ -C_3 |\lambda_2(t)|^{-3/4} M^{1/4} \times \operatorname{sgn}[M(t)], & \text{if } |\sigma| < \sigma_{max}. \end{cases}$$

For a given deflection requirement Δ there are three possibilities concerning attainment of the maxium deflection Either

1. $|x(t)| < \Delta$ for all t,

2. $|x(t)| = \Delta$ for just one t, or

3. $|x(t)| = \Delta$ for two distinct t.

The third possibility occurs here because $M(t)$ changes sign. In Case 1, $|\sigma| = \sigma_{max}$ throughout the beam determines $h(t)$. The Case 2 may be treated in exactly the same way as the problem in the preceding paragraph. Case 3 is considered in detail here.

Assume there are two points at which

$|x(t)| = \Delta$; the beam will not probably be subdivided as shown in Fig. 7-15.

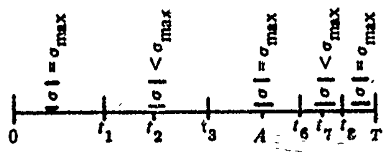


Figure 7-15. Subdivision of the Beam

Points t_2 and t_7, where Case 3 occurs, are t_r^-, so they are determined by Eq. 7-102. Further, $x_1(t_2) = \Delta$ and $x_1(t_7) = -\Delta$. Points t_1, t_3, t_6, and t_8 are t_η^* and are determined by Eq. 7-107.

In this problem, t_2 and t_7 may be points of discontinuity of λ_1 and λ_2. The conditions, Eqs. 7-85 and 7-86 on λ_2 are $\lambda_2(0) = \lambda_2(T) = 0$. Therefore, λ_2 may be written as

$$\lambda_2(t) = \begin{cases} \zeta_1 t & \text{, if } 0 < t < t_2 \\ \zeta_3 - \zeta_2 t & \text{, if } t_2 < t < t_7 \\ \zeta_4\left(\dfrac{t}{T} - 1\right) & \text{, if } t_7 < t < t. \end{cases}$$

Solving Eqs. 7-102 and 7-107 for the t_i yields the following:

$$t_1 = \left(\frac{EM}{\sigma^2_{max}}\right)\frac{1}{\zeta_1} \tag{7-131}$$

$$t_2 = \frac{\zeta_3}{\zeta_1 + \zeta_2} \tag{7-132}$$

$$t_3 = \frac{\zeta_3}{\zeta_2} - \left(\frac{EM}{\sigma^2_{max}}\right)\frac{1}{\zeta_2} \tag{7-133}$$

$$t_6 = \frac{\zeta_3}{\zeta_2} + \left(\frac{EM}{\sigma^2_{max}}\right)\frac{1}{\zeta_2} \tag{7-134}$$

$$t_7 = \frac{\zeta_3 + \zeta_4}{\dfrac{\zeta_4}{T} + \zeta_2} \tag{7-135}$$

and

$$t_8 = T\left[1 - \left(\frac{EM}{\sigma^2_{max}}\right)\frac{1}{\zeta_4}\right]. \tag{7-136}$$

In obtaining the expressions for t_3 and t_6, use was made of the fact that $t_3 < A < t_6$. The result, $\lambda_2(t)M(t) > 0$, from Eqs. 7-97 and 7-98 shows that this is true. To prove this, assume for definiteness $t_3 > A$. Since $M(t)$ changes sign at A and $\lambda_2(t)$ is continuous there, $\lambda_2(A) = 0$. But, since $\lambda_2(t)$ is continuous near A, it is arbitrarily near zero in a neighborhood of A. Eq. 7-130 then shows that $h(t)$ is arbitrarily near zero in a neighborhood of A and this violates the condition $|\sigma| < \sigma_{max}$. Likewise, $A < t_6$.

Conditions that determine the ζ_i are

$$x_1(t_2) = \Delta, x_2(t_2) = 0$$

and

$$x_1(t_7) = -\Delta, x_2(t_7) = 0.$$

For computational reasons, it is more convenient to use the following equivalent set of conditions:

$$x_1(0) = 0$$

$$x_2(t_7) = 0$$

$$x_1(t_7) + \Delta = 0$$

and

$$x_1(T) = 0$$

where $x_1(t_2) = \Delta$ and $x_2(t_2) = 0$ are used as initial conditions for integration.

More explicity, these equations are

$$R_1 = x_1(0) = \Delta + t_1 \int_0^{t_1} f_2(\eta)d\eta$$

$$+ t_2 \int_{t_1}^{t_2} f_2(\eta,\zeta_1)d\eta$$

$$+ \int_0^{t_1} \int_0^{\nu} f_2(\eta)d\eta d\nu$$

$$- \int_{t_1}^{t_2} \int_{t_1}^{\nu} f_2(\eta,\zeta_1)d\eta d\nu = 0 \tag{7-137}$$

$$R_2 = x_2(t_7) = \int_{t_2}^{t_3} f_2(\eta,\zeta_2,\zeta_3)d\eta$$

$$+ \int_{t_3}^{t_4} f_2(\eta)d\eta$$

$$+ \int_{t_4}^{t_7} f_2(\eta,\zeta_2,\zeta_3)d\eta = 0 \tag{7-138}$$

$$R_3 = x_1(t_7) + \Delta = 2\Delta$$

$$+ (t_7 - t_3)\int_{t_2}^{t_3} f_2(\eta,\zeta_2,\zeta_3)d\eta$$

$$+ (t_7 - t_6)\int_{t_3}^{t_4} f_2(\eta)d\eta$$

$$+ \int_{t_2}^{t_3} \int_{t_2}^{\nu} f_2(\eta,\zeta_2,\zeta_3)d\eta d\nu$$

$$+ \int_{t_3}^{t_4} \int_{t_3}^{\nu} f_2(\eta)d\eta d\nu$$

$$+ \int_{t_4}^{t_7} \int_{t_4}^{\nu} f_2(\eta,\zeta_2,\zeta_3)d\eta d\nu = 0 \tag{7-139}$$

and

$$R_4 = x(T) = -\Delta$$

$$+ (T - t_8) \int_{t_7}^{t_8} f_2(\eta,\zeta_4)d\eta$$

$$+ \int_{t_7}^{t_8} \int_{t_7}^{\nu} f_2(\eta,\zeta_4)d\eta d\nu$$

$$+ \int_{t_8}^{T} \int_{t_8}^{\nu} f_2(\eta)\, d\eta d\nu = 0 \tag{7-140}$$

where the R_i are introduced for notational purposes.

Generalized Newton Method is used to solve Eqs. 7-137 through 7-140 for the ζ_i. An initial estimate $\hat{\zeta}_i$, $i = 1, 2, 3, 4$, is made; and a correction is computed according to the formula

$$\begin{bmatrix} \bar{\xi}_1 \\ \bar{\xi}_2 \\ \bar{\xi}_3 \\ \bar{\xi}_4 \end{bmatrix} = \begin{bmatrix} \hat{\xi}_1 \\ \hat{\xi}_2 \\ \hat{\xi}_3 \\ \hat{\xi}_4 \end{bmatrix} - \left[\frac{\partial R_i}{\partial \xi_j}\right]^{-1} \begin{bmatrix} R_1 \\ R_2 \\ R_3 \\ R_4 \end{bmatrix} \tag{7-141}$$

where terms on the right side of Eq. 7-141 are computed in terms of the $\hat{\xi}_i$. The corrected guess $\bar{\xi}_i$ then takes the place of $\hat{\xi}_i$ and the process is repeated.

The derivatives of the R_i with respect to the ξ_i are computed by the chain rule of differentiation, Eq. 7-109, and Eqs. 7-137 through 7-140. Just as in Eq. 7-129 several of these derivatives must be determined by successive numerical integration.

The matrix of derivatives which appears in Eq. 7-141 has sixteen elements. Twelve successive definite integrals appear in one or more elements of this matrix. Therefore, considerable computation is involved in each iteration. All this computation was incorporated in a single computer program.

As a numerical example, the beam of Fig. 7-14 having the following properties is considered:

M = 1,100 in.-lb

T = 40 in.

b = 0.5 in.

E = 10^7 lb/in.^{2}

A = 25 in.

τ_{max} = 10,000 lb/in.^{2}

and

σ_{max} = 20,000 lb/in.^{2}

It was noted that three distinct situations may occur depending on the value of Δ. In this example, the problem breaks down as follows:

1. If $\Delta > 0.509$, then $x_1(t) < \Delta$ for all t,

2. If $0.156 < \Delta < 0.509$, then there is just one point t for which $|x_1(t)| = \Delta$, and

3. If $\Delta < 0.156$, then there are two values of t for which $|x_1(t)| = \Delta$.

The numerical example given was solved for eleven different values of Δ. The results of these calculations are presented in Table 7-5. The first column of this table consists of the values of Δ considered. The following ten columns give information that, when substituted into Eq. 7-130, completely specifies the optimal beam. The next column gives the volume of this optimal beam. The final two columns give the volume of the lightest beam constant depth that satisfies the conditions of the problem and the percent saving realized when the optimal beam is used instead of this uniform beam. Dashes have been inserted in the table when the quantity to be tabulated does not exist.

For each value of Δ, the iterative procedure used to solve the problem required approximately two minutes per iteration on an IBM 1410 Computer. However, three to five iterations were sufficient to obtain convergence of the residue to seven decimal places, so computing time was not excessive. This rapid convergence is, again, characteristic of the Generalized Newton Method.

An interesting sidelight of this particular

TABLE 7-5
RESULTS FOR BEAM WITH S-SHAPED DEFLECTION CURVE

Δ, in.	t_1, in.	t_2, in.	t_3, in.	t_6, in.	t_7, in.	t_8, in.	ξ_1	ξ_2	ξ_3	ξ_4	Vol., in.^{3}	Vol. of Unif. Beam[1], in.^{3}	Savings, %
>0.509	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	16.3	16.3	0
0.40	8.27	13.59	23.93	--------	--------	--------	3.35	--------	--------	69.00	16.7	17.6	5
0.30	5.56	12.75	25.00	--------	--------	--------	4.94	--------	--------	92.91	17.4	19.4	12
0.23	4.18	11.94	25.00	--------	--------	--------	6.57	--------	--------	111.96	17.9	21.1	18
0.16	2.84	10.96	25.00	--------	--------	--------	9.69	--------	--------	146.37	18.8	23.9	28
0.15	2.32	11.56	24.51	31.02	32.02	33.90	11.84	6.45	234.51	180.41	19.0	24.3	29
0.14	2.11	11.54	24.56	29.66	31.93	35.09	13.10	9.75	263.99	225.17	19.3	25.0	30
0.10	1.31	11.53	24.58	27.93	31.77	37.50	21.02	16.45	431.98	440.45	21.3	28.0	31
0.075	0.88	11.49	24.74	26.95	31.67	38.42	31.10	24.89	643.33	695.92	23.3	30.7	32
0.05	0.51	11.45	24.89	26.15	31.58	39.12	53.56	43.62	1113.10	1256.79	26.5	35.2	33
0.02	0.15	11.40	25.00	25.37	31.49	39.75	182.14	150.63	3794.73	4457.51	35.9	47.7	33

[1] Volume of lightest beam of constant depth which satisfies all the requirements of the problem.

example concludes the present subparagraph. A plot of volume (of optimal beam) versus deflection requirement for the problem considered here is given in Fig. 7-16. From this

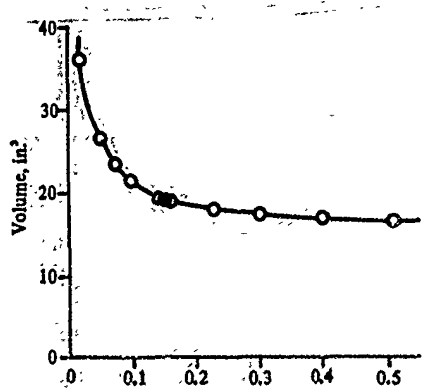


Figure 7-16. Volume vs Deflection Requirement

graph, it appears that the volume of the optimal beam is a continuous function of deflection requirement. This is a rather remarkable result in view of the fact that optimal beams with deflection requirements greater and less than 0.156 have considerably different form. The beam profiles of Figs. 7-17 and 7-18 illustrate this difference graphically. As Δ decreases toward 0.156, the jump in $h(t)$ at $t = 25$ (see Fig. 7-17) becomes more pronounced. However, for Δ very slightly less than 0.156, the profile is continuous, much as in Fig. 7-18.

7-4.5.4 CONCLUSIONS

The examples considered in pars. 7-4.5.2 and 7-4.5.3 are of the order of complexity that might be found in actual practice. In these examples, a saving of material up to 33% is realized when nonuniform, optimal beams are used instead of uniform beams. For more complex loading situations, the saving may be even greater. From an engineering viewpoint, such savings are significant.

In structural applications, this saving may be offset by additional cost of fabrication.

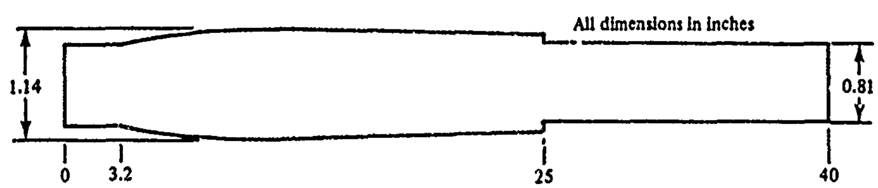


Figure 7-17. Profile of Optimal Beam for Δ = 0.16

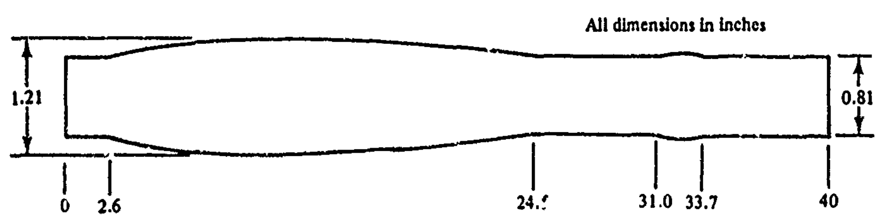


Figure 7-18. Profile of Optimal Beam for Δ = 0.15

However, for applications in which weight is a premium, such as in aerospace work, fabrication of minimum weight structural members may be quite feasible. Further, if the cost of forming nonuniform beams is not prohibitive, such as in the manufacture of reinforced concrete beams, then nonuniform optimal beams may be used to advantage.

REFERENCES

1. R. A. Ridha and R. N. Wright, "Minimum Cost Design of Frames", *J. of the Structural Division*, ASCE, Vol. 93, No. ST4, pp. 165-183, August 1967.
2. J.L. Lagrange, "Sur la forces des Ressorts plies", *Mem. Acad.*, Berlin, 1771.
3. T. Clausen, "Uber die For architektonischer Saulen", *Bulletin physicomethematique de l'Academie*, T. I X., pp. 386-379, St. Petersburg, 1851.
4. Z. Wasiutynski, and A. Brandt, "The Present State of Knowledge in the Field of Optimum Design of Structures", *App. Mech. Reviews*, Vol. 16, No. 5, pp. 341-350, May 1963.
5. C.Y. Sheu and W. Prager, "Recent Developments in Optimal Structural Design", *App. Mech. Reviews*, Vol. 21, No. 10, pp. 985-992, October 1968.
6. D.C. Drucker and R.T. Shield, "Design for Minimum Weight", *Proc. 9th International Congr. Appl. Mech.*, Brussels, 1956.
7. W. Prager, "Minimum-Weight Design of a Portal Frame", *Proc. ASCE* Vol. 82, 1956.
8. D.C. Drucker, "On Minimum Weight Design and Strength of Nonhomogeneous Plastic Bodies", *Proc. Symposium IUTAM*, pp. 139-146, Warsaw, 1958.
9. J.B.Keller, "The Shape of the Strongest Column", *Archive Ratl. Mech. Anal.*, Vol. 5, pp. 275-285, 1960.
10. I. Tadjbakhsh and J.B. Keller, "Strongest Columns and Isoperametric Inequalities for Eigenvalues", *J. Appl. Mech.*, Vol. 29, No. 1, pp. 159-164, March 1962.
11. J.B. Keller and F.I. Niordson, "The Tallest Column", *J. of Math. and Mech.*, Vol. 16, No. 5, pp. 433-446, 1966.
12. F.I. Niordson, "The Optimal Design of a Vibration Beam", *Quarterly of Appl. Math.*, Vol. 23, No. 1, pp. 47-53, April 1965.
13. W. Prager and J.E. Taylor, "Problems of Optimal Structural Design", *J. Appl. Mech.*, Vol. 35, No. 1, pp. 102-106, March 1966.
14. J.E. Taylor, "Minimum Mass Bar for Axial Vibration at Specified Natural Frequency", *AIAA J.*, Vol. 5, No. 10, pp. 1911-1913, October 1967.
15. J.E. Taylor and C.Y. Liu, "On the Optimal Design of Columns", *AIAA J.*, Vol. 6, No. 8, August 1968.

16. C. Goffman, *Calculus of Several Variables*, Harper and Row, New York, 1965.

17. G.A. Bliss, *Lectures on the Calculus of Variations*, University of Chicago Press, Chicago, 1946.

18. J.F. Traub, *Iterative Methods for the Solution of Equations*, Prentice-Hall, Englewood Cliffs, New Jersey, 1964.

CHAPTER 8

METHODS OF STEEPEST DESCENT FOR OPTIMAL DESIGN PROBLEMS

8-1 INTRODUCTION

As seen by the examples of Chapter 7, solution of the necessary conditions for the general problem of optimal design is difficult. Even in idealized design problems numerical methods must normally be employed to construct a solution.

The numerical techniques for the indirect method presented in par. 6-5 and in Chapter 7 are iterative in nature. Each of the techniques requires that an estimate of the solution be made before the iterative process may be initiated. In many cases, particularly in new problem areas, the designer may have only a gross notion of what to expect of the solution so his initial estimate may be poor.

Convergence of the techniques of Chapters 6 and 7 are reported to be very poor unless good estimates of the solution are available. In fact, these iterative techniques often diverge for poor estimates of the solution. On the other hand, if a good initial estimate is available, these methods converge very rapidly.

This discussion illustrates the need for a workhorse technique that may be used even when only poor estimates of the solution of the optimal design problem are available. The method should be capable of making steady improvement in an estimated solution and, in fact, converge to the solution. Rate of convergence could be sacrificed for dependability if required.

A second desirable property of a general method of optimal design is that it apply routinely to a large class of real-world optimal design problems. To be useful to the working design engineer, the method should apply whenever the designer has developed the capability to analyze the system to be designed. Further, the method should be explicit enough so that a senior engineer can set the problem up for computation and a less experienced junior engineer can program the algorithm for use on a digital computer.

The methods to be developed in this chapter and applied in the next have many of these nice-to-have properties. The basic idea of these direct methods is to simplify the basic design problem so that it will readily yield information which allows the designer to make a small improvement in an estimated optimum design. After the improvement is made, a new and better estimate of the solution of the optimal design problem is obtained. The process is repeated successively to obtain small improvements in the best available estimate of the solution until the design obtained is sufficiently near the optimum.

The basic method of simplification of the design problem is to expand functions involved in the problems through use of Taylor's Formula. In this way, a simplified problem is obtained which serves as a good approximation of the original problem provided only small changes are allowed in certain variables.

8-2 A STEEPEST DESCENT METHOD FOR THE BASIC OPTIMAL DESIGN PROBLEM

8-2.1 THE PROBLEM CONSIDERED

In order to present the basic ideas of the Method of Steepest Descent, consideration here will be limited to optimal design problems with fixed endpoints, no discontinuities in the basic problems, and no intermediate conditions on the state variable. As seen in par. 6-4, this eliminates state variable inequality constraints from direct treatment. All these features of more general optimal design problems will be treated in par. 8-3.

Specifically, the problem treated here is to find $u(t)$, $t^0 < t < t^1$, and b which minimize

$$J = g_0(b,x^0,x^1) + \int_{t^0}^{t^1} f_0[t,x(t),u(t),b]\,dt \tag{8-1}$$

subject to the conditions

$$\left.\begin{aligned} &\frac{dx}{dt} = f(t,x,u,b), && t^0 < t < t^1 \\ &\theta_s(x^0,x^1) = 0 && s = 1, ..., n \end{aligned}\right\} \tag{8-2}$$

$$\left.\begin{aligned} &\psi_\alpha = g_\alpha(b,x^0,x^1) + \int_{t^0}^{t^1} L_\alpha[t,x(t),u(t),b]\,dt = 0, \quad \alpha = 1, ..., r' \\ &\psi_\alpha = g_\alpha(b,x^0,x^1) + \int_{t^0}^{t^1} L_\alpha[t,x(t),u(t),b]\,dt < 0, \quad \alpha = r' + 1, ..., r \end{aligned}\right\} \tag{8-3}$$

and

$$\left.\begin{aligned} &\phi_\beta(t,u) = 0, \beta = 1, ..., q', \ t^0 < t < t^1, \\ &\phi_\beta(t,u) < 0, \ \beta = q' + 1, ..., q, \ t^0 < t < t^1. \end{aligned}\right\} \tag{8-4}$$

Just as in Chapter 6, the variables $x(t)$, $u(t)$, and b are vectors, $x(t) = [x_1(t), ..., x_n(t)]^T$, $u(t) = [u_1(t), ..., u_m(t)]^T$, and $b = [b_1, ..., b_k]^T$.

Any inequality constraints of the form

$$\omega(t,x,u,b) < 0 \tag{8-5}$$

can be transformed into a constraint of the form

$$\int_{t^0}^{t^1} \left\{\omega[t,x(t),u(t),b] + |\,\omega[t,x(t),u(t),b]\,|\right\} dt = 0 \tag{8-6}$$

which is then a constraint of the kind of Eq. 8-3.

The class of problems considered is, therefore, fairly general. The essential features that are not included are variable limits of integration, discontinuities in functions of the problem ($u(t)$ may still be discontinuous), intermediate conditions on $x(t)$, and state variable inequality constraints.

8-2.2 EFFECTS OF SMALL CHANGES IN DESIGN VARIABLES AND PARAMETERS

The basic idea of the direct method of solving optimal design problems is to first construct an estimate $u^{(0)}(t)$, $b^{(0)}$ of the solution and then find small changes $\delta u(t)$, δb such that $u^{(0)}(t) + \delta u(t)$, $b^{(0)} + \delta b$ is an improved estimate in some sense. Before the improvements can be determined, analysis of their effect on the problem must be performed.

In this analysis, $\delta u(t)$ and δb are required to be small so that a first order Taylor expansion of the functions of the problem is a good approximation. Since $x(t)$ is the solution of a boundary-value problem involving $u(t)$ and b, it is clear that $\delta u(t)$, δb will cause a change $\delta x(t)$ in $x(t)$. It is assumed here that the boundary-value problem for $x(t)$ is well posed, see Ref. 1, page 227, so that $\delta u(t)$, δb small implies $\delta x(t)$ small. Using this fact,

$$\left.\begin{aligned} \delta J = {} & \frac{\partial g_0}{\partial b}\,\delta b + \frac{\partial g_0}{\partial x^0}\,\delta x^0 + \frac{\partial g_0}{\partial x^1}\,\delta x^1 \\ & + \int_{t^0}^{t^1}\left(\frac{\partial f_0}{\partial x}\,\delta x + \frac{\partial f_0}{\partial u}\,\delta u \right. \\ & \left. + \frac{\partial f_0}{\partial b}\,\delta b\right) dt \end{aligned}\right\} \quad (8\text{-}7)$$

$$\left.\begin{aligned} & \frac{d\delta x}{dt} = \frac{\partial f}{\partial x}\,\delta x + \frac{\partial f}{\partial u}\,\delta u + \frac{\partial f}{\partial b}\,\delta b \\ & \frac{\partial \theta_s}{\partial x^0}\,\delta x^0 + \frac{\partial \theta_s}{\partial x^1}\,\delta x^1 = 0,\ s = 1, \ldots, n \end{aligned}\right\} \quad (8\text{-}8)$$

$$\begin{aligned} \delta \psi_\alpha = {} & \frac{\partial g_\alpha}{\partial b}\,\delta b + \frac{\partial g_\alpha}{\partial x^0}\,\delta x^0 + \frac{\partial g_\alpha}{\partial x^1}\,\delta x^1 \\ & + \int_{t^0}^{t^1}\left(\frac{\partial L_\alpha}{\partial x}\,\delta x + \frac{\partial L_\alpha}{\partial u}\,\delta u \right. \\ & \left. + \frac{\partial L_\alpha}{\partial b}\,\delta b\right) dt \end{aligned} \quad (8\text{-}9)$$

$\alpha = 1, \ldots, r$, and

$$\delta \phi_\beta = \frac{\partial \phi_\beta}{\partial u}\,\delta u,\ \beta = 1, \ldots, q\,. \quad (8\text{-}10)$$

In all the formulas, Eqs. 8-7 through 8-10, the functions are evaluated at $[t, x^{(0)}(t), u^{(0)}(t), b^{(0)}]$ where $x^{(0)}(t)$ is the solution of the boundary-value problem Eq. 8-2 for $x(t)$ with $u(t) = u^{(0)}(t)$ and $b = b^{(0)}$.

To simplify the work which follows, it will be convenient to eliminate explicit dependence of Eqs. 8-7, 8-9, and 8-10 on $\delta x(t)$. This elimination is performed through use of the differential equation adjoint to the linear equation for $\delta x(t)$ in Eq. 8-8 (Ref. 2). This equation is

$$\frac{d\lambda}{dt} = -\frac{\partial f^T}{\partial x}\,\lambda + h(t) \quad (8\text{-}11)$$

where the function $h(t)$ will be chosen to obtain results needed later in the development.

Note that for any solution $\lambda(t)$ of Eq. 8-11 and any solution $\delta x(t)$ of Eq. 8-8,

$$\begin{aligned} \frac{d}{dt}(\lambda^T \delta x) = {} & \frac{d\lambda^T}{dt}\,\delta x + \lambda^T \frac{d\delta x}{dt} = -\lambda^T \frac{\partial f}{\partial x}\,\delta x \\ & + h^T \delta x + \lambda^T \frac{\partial f}{\partial x}\,\delta x + \lambda^T \frac{\partial f}{\partial u}\,\delta u \\ & + \lambda^T \frac{\partial f}{\partial b}\,\delta b = h^T \delta x + \lambda^T \frac{\partial f}{\partial u}\,\delta u \\ & + \lambda^T \frac{\partial f}{\partial b}\,\delta b. \end{aligned}$$

Integrating this equation from t^0 to t^1 and using the fundamental theorem of calculus,

$$\lambda^T(t^1)\delta x(t^1) - \lambda^T(t^0)\delta x(t^0) =$$

$$\int_{t^0}^{t^1} \left(h^T \delta x + \lambda^T \frac{\partial f}{\partial u} \delta u + \lambda^T \frac{\partial f}{\partial b} \delta b \right) dt. \tag{8-12}$$

By choosing the function $h(t)$ and the boundary conditions on $\lambda(t)$ appropriately, the identity Eq. 8-12 will yield the desired relationships. First, put $h(t) = -\partial f_0{}^T/\partial x$ $[t,u^{(0)},b^{(0)}]$ and define $\lambda^J(t)$ as the solution of Eq. 8-11 with boundary conditions on $\lambda^J(t^0)$ and $\lambda^J(t^1)$ determined by

$$\lambda^{J^T}(t^1)\delta x^1 - \lambda^{J^T}(t^0)\delta x^0$$

$$= \frac{\partial g_0}{\partial x^0} \delta x^0 + \frac{\partial g_1}{\partial x^1} \delta x^1 \tag{8-13}$$

for all δx^0 and δx^1 satisfying the second equation of Eq. 8-8. To see that the second equation, Eq. 8-8, and Eq. 8-13 determine conditions on $\lambda^J(t^0)$ and $\lambda^J(t^1)$, consider the following procedure. Determine n of the $2n$ variables $\delta x^0{}_i$, $\delta x^1{}_i$, $i = 1, ..., n$ in terms of the remaining n of these variables. Now substitute the variables $\delta x^0{}_i$, $\delta x^1{}_i$ just found into Eq. 8-13. Eq. 8-13 may now be written as a linear combination of n independent $\delta x^0{}_i$, $\delta x^1{}_i$. Since Eq. 8-13 must hold for all n independent variables $\delta x^0{}_i$, $\delta x^1{}_i$ previously identified, the coefficients of all these variables must be zero. This is then a system of n equations involving only $\lambda^J(t^0)$, $\lambda^J(t^1)$ and known quantities. This procedure will be carried out in detail in particular problems

Substituting from Eq. 8-12 into Eq. 8-7 yields

$$\delta J = \frac{\partial g_0}{\partial b} \delta b + \int_{t^0}^{t^1} \left[\left(\frac{\partial f_0}{\partial b} + \lambda^{J^T} \frac{\partial f}{\partial u} \right) \delta u + \left(\frac{\partial f_0}{\partial b} + \lambda^{J^T} \frac{\partial f}{\partial b} \right) \delta b \right] dt. \tag{8-14}$$

Likewise, put $h(t) = -\frac{\partial L_\alpha}{\partial x}[t,u^{(0)},b^{(0)}]$ and define $\lambda^{\psi_\alpha}(t)$ as the solution of Eq. 8-11 with boundary conditions on $\lambda^{\psi_\alpha}(t^0)$ and $\lambda^{\psi_\alpha}(t^1)$ determined by

$$\lambda^{\psi_\alpha{}^T}(t^1)\delta x^1 - \lambda^{\psi_\alpha{}^T}(t^0)\delta x^0$$

$$= \frac{\partial g_\alpha}{\partial x^0} \delta x^0 + \frac{\partial g_\alpha}{\partial x^1} \delta x^1 \tag{8-15}$$

for all δx^0 and δx^1 satisfying the second equation of Eq. 8-8. The identity Eq. 8-15 determined boundary conditions just as Eq. 8-13 did. Substituting from Eq. 8-12 into Eq. 8-9 yields

$$\delta \psi_\alpha = \frac{\partial g_\alpha}{\partial b} \delta b + \int_{t^0}^{t^1} \left[\left(\frac{\partial L_\alpha}{\partial u} + \lambda^{\psi_\alpha{}^T} \frac{\partial f}{\partial u} \right) \delta u + \left(\frac{\partial L_\alpha}{\partial b} + \lambda^{\psi_\alpha{}^T} \frac{\partial f}{\partial b} \right) \delta b \right] dt. \tag{8-16}$$

In terms of the adjoint variables $\lambda^J(t)$ and $\lambda^{\psi_\alpha}(t)$, the quantities δJ and $\delta \psi_\alpha$ are now given explicitly as functions of $\delta u(t)$ and δb. The problem is now reduced to determining $\delta u(t)$ and δb, which yield the greatest reduction in J subject to the linearized constraints of the problem.

It should be noted that the boundary-value problems for $\lambda^J(t)$ and $\lambda^{\psi_\alpha}(t)$, $\alpha = 1, ..., r$, have solutions if the boundary-value problem of Eq. 8-8 is well posed This is a basic property of adjoint boundary-value problems which is proved in Ref. 2.

8-2.3 A STEEPEST DESCENT APPROACH

The problem of determining $\delta u(t)$ and δb to minimize (maximum negative of) δJ must deal with the inequality constraints, Eqs. 8-3 and 8-4. The argument to be used here is: simply ignore a constraint function that is negative before an iteration begins. If, on the other hand, a constraint function is positive, it is required to be reduced. For example, if $\psi_\alpha > 0$ or $\phi_\beta(t) > 0$ for some t, then it is required that $\delta\psi_\alpha = -a\psi_\alpha$, $\alpha = 1, ..., r'$ and $\delta\psi_\alpha < -a\psi_\alpha$, $\alpha = r' + 1, ..., r$ and $\psi_\alpha > 0$ and $\delta\phi_\beta(t) = -c\phi_\beta(t)$, $\beta = 1, ..., q'$ and $\delta\phi_\beta(t) < -c\phi_\beta(t)$, $\beta = q' + 1, ..., q$ and $\phi_\beta(t) > 0$ where $0 < a < 1$ and $0 < c < 1$. The magnitude of a and c are chosen so that the required changes $\delta\psi_\alpha$ and $\delta\phi_\beta(t)$ are not excessively large. If ψ_α and $\phi_\beta(t)$ are not so large that the linear approximation is violated with $a = 1$, or $c = 1$, then a or c are chosen as one.

For convenience, define two sets of indices

$$A = \left\{ \alpha \mid \psi_\alpha[x^{(0)}, u^{(0)}, b^{(0)}] > 0 \right\}$$

and

$$B(t) = \left\{ \beta \mid \phi_\beta[t, u^{(0)}(t)] > 0 \right\}.$$

It should be noted that the collection $B(t)$ of indices may change with the variable t.

Define the column vector of elements ψ_α with $\psi_\alpha > 0$

$$\tilde{\psi} = \begin{bmatrix} \psi_\alpha \\ \\ \alpha \in A \end{bmatrix} \tag{8-17}$$

and a similar column vector of functions $\phi_\beta(t)$ with $\phi_\beta(t) > 0$

$$\tilde{\phi}(t) = \begin{bmatrix} \phi_\beta(t) \\ \\ \beta \in B(t) \end{bmatrix}. \tag{8-18}$$

Note that the column vector $\tilde{\phi}(t)$ may have different components at different points in $t^0 < t < t^1$. In order to assure that constraints are satisfied, it will be required that

$$\left.\begin{aligned} &\delta\psi_\alpha = -a\psi_\alpha \quad \alpha = 1, ..., r' \\ &\delta\psi_\alpha < -a\psi_\alpha, \; \alpha = r' + 1, ..., r \\ &\qquad \text{and } \alpha \in A \end{aligned}\right\} \tag{8-19}$$

and

$$\left.\begin{aligned} &\delta\phi_\beta(t) = -c\phi_\beta(t), \quad \beta = 1, ..., q' \\ &\delta\phi_\beta(t) < -c\phi_\beta(t), \quad \beta = q' + 1, ..., q \\ &\qquad \text{and } \beta \in B(t). \end{aligned}\right\} \tag{8-20}$$

Finally, define

$$\Lambda^J(t) = \frac{\partial f_0{}^T}{\partial u} + \frac{\partial f^T}{\partial u} \lambda^J(t) \tag{8-21}$$

$$\ell^J = \frac{\partial g_0{}^T}{\partial b} + \int_{t^0}^{t^1} \left[\frac{\partial f_0{}^T}{\partial b} + \frac{\partial f^T}{\partial b} \lambda^J(t) \right] dt \tag{8-22}$$

$$\Lambda^\psi(t) = \left[\frac{\partial L_\alpha{}^T}{\partial u} + \frac{\partial f^T}{\partial u} \lambda^{\psi_\alpha} \right], \quad \text{for all } \alpha \in A \tag{8-23}$$

and

$$\ell^\psi = \left[\frac{\partial g_\alpha{}^T}{\partial b} + \int_{t^0}^{t^1} \left(\frac{\partial L_\alpha{}^T}{\partial b} + \frac{\partial f^T}{\partial b} \lambda^{\psi_\alpha} \right) dt \right], \quad \text{for all } \alpha \in A. \tag{8-24}$$

Note that $\Lambda^\psi(t)$ is a matrix of functions with m rows and the same number of columns as there are indices in A. The matrix ℓ^ψ of constants has k rows and the same number of columns as $\Lambda^{\psi}(t)$.

Using the matrix notation of Eqs 8-21 through 8-24 in Eqs. 8-14, 8-16, and 8-17.

$$\delta J = \ell^{J\,T}\delta b + \int_{t^0}^{t^1} \Lambda^{J\,T}(t)\delta u\,dt \tag{8-25}$$

$$\delta\tilde{\psi} = \ell^{\psi\,T}\delta b + \int_{t^0}^{t^1} \Lambda^{\psi\,T}(t)\delta u\,dt\,. \tag{8-26}$$

Before $\delta u(t)$ and δb are determined, some mechanism must be set up for requiring that these variations are actually small. For convenience, put

$$dP^2 = \delta b^T W_b \delta b + \int_{t^0}^{t^1} \delta u^T W_u(t)\delta u\,dt \tag{8-27}$$

where W_b and $W_u(t)$ are chosen as positive definite weighting matrices and dP is to be chosen small enough that $\delta u(t)$ and δb are sufficiently small.

The problem is now reduced to finding $\delta u(t)$ and δb which minimize δJ subject to Eqs. 8-19, 8-20, and 8-27. This problem is now a special case of the Bolza problem of par. 6-4. According to Theorem 6-7, there exist multipliers $\lambda_0 > 0$, $\gamma = \begin{bmatrix}\gamma_\alpha \\ \alpha\in A\end{bmatrix}$, $\gamma_\alpha > 0$ for $\alpha > r'$, $\mu(t) = \begin{bmatrix}\mu_\beta(t) \\ \beta\in B(t)\end{bmatrix}$ $\mu_\beta(t) > 0$ for $\beta > q'$, and γ_0 with

$$H = \left[-\lambda_0\Lambda^{J\,T} - \gamma^T\Lambda^{\psi\,T}(t) - \mu^T(t)\frac{\partial\tilde{\phi}}{\partial u}\right]\delta u - \gamma_0\delta u^T W_u\delta u + \mu^T c\tilde{\phi} \tag{8-28}$$

$$G = \left(\lambda_0\Lambda^{J\,T} + \gamma^T\ell^{\psi\,T}\right)\delta b + \gamma_0\delta b^T W_b\delta b + \gamma^T c\,\tilde{\psi} \tag{8-29}$$

such that

$$\frac{\partial H}{\partial\delta u} = 0 = -\lambda_0\Lambda^{J\,T} - \gamma^T\Lambda^{\psi\,T} - \mu^T(t)\frac{\partial\tilde{\phi}}{\partial u} - 2\gamma_0\delta u^T W_u \tag{8-30}$$

$$\frac{\partial G}{\partial\delta b} - \int_{t^0}^{t^1}\frac{\partial H}{\partial\delta b}\,dt = 0 = \lambda_0\ell^{J\,T} + \gamma^T\ell^{\psi\,T} + 2\gamma_0\delta b^T W_b\,. \tag{8-31}$$

In the following development, it will be assumed that the problem is normal so that it is permissible to put $\lambda_0 = 1$. Solving Eqs. 8-30 and 8-31 for $\delta u(t)$ and δb, respectively, yields

$$\delta u(t) = -\frac{1}{2\gamma_0}W_u(t)^{-1} \times\left[\Lambda^J(t) + \Lambda^\psi(t)\gamma + \frac{\partial\tilde{\phi}^T}{\partial u}\mu(t)\right] \tag{8-32}$$

and

$$\delta b = -\frac{1}{2\gamma_0}W_b^{-1}\left(\ell^J + \ell^\psi\gamma\right) \tag{8-33}$$

In order to complete the determination of $\delta u(t)$ and δb, $\mu(t)$ and γ must be eliminated from Eqs. 8-32 and 8-33. A direct analytical elimination of $\mu(t)$ and γ is not feasible at this point since Theorem 6-7 requires

$$\gamma^T(\delta\bar{\psi} + a\bar{\psi}) = 0$$

$$\mu^T(t)[\delta\bar{\phi}(t) + c\bar{\phi}(t)] = 0$$

and some components of $\delta\bar{\psi} + a\bar{\psi}$ and $\delta\bar{\phi}(t) + c\bar{\phi}(t)$ will be zero. A certain amount of logic is required to find γ and $\mu(t)$. Since only small $\delta u(t)$ and δb are admitted, if $\bar{\psi}$ and $\bar{\phi}(t)$ are zero for $u^{(0)}(t)$, $b^{(0)}$ then very likely they will also be zero for $u^{(0)}(t) + \delta u(t)$, $b^{(0)} + \delta b$. Following this line of reasoning, it will be assumed first that

$$\delta\bar{\psi} + a\bar{\psi} = 0$$

and

$$\delta\bar{\phi}(t) + c\bar{\phi}(t) = 0\,.$$

Then γ and $\mu(t)$ are determined by substituting $\delta u(t)$ and δb from Eqs. 8-32 and 8-33 into these equations. The multipliers γ and $\mu(t)$ are then determined and checked for the proper sign. If $\gamma_\alpha > 0$ for $\alpha > r'$ and $\mu_\beta(t) > 0$ for $\beta > q'$, then this assumption is admissible. If, on the other hand, $\gamma_\alpha < 0$ for some $\alpha > r'$, then $\delta\psi_\alpha + a\psi_\alpha = 0$ is incorrect and it must be that $\delta\psi_\alpha + a\psi_\alpha < 0$ should occur. This is equivalent to simply removing ψ_α from $\bar{\psi}$ and recalculating. Likewise, if $\mu_\beta(t) < 0$ for some $\beta > q'$, then $\delta\phi_\beta(t) + a\phi_\beta(t) < 0$ should occur and $\phi_\beta(t)$ should be removed from $\bar{\phi}(t)$ and the multipliers recalculated. So much for the semi-mathematics, now to the calculations based on this argument.

If $B(t)$ is empty for all t, then $\bar{\phi}$ is not defined and $\mu(t)$ need not be determined. In case $B(t)$ is not empty, is to be required that

$$\delta\bar{\phi} + c\bar{\phi} = 0 = \frac{\partial\bar{\phi}}{\partial u}\,\delta u + c\bar{\phi}$$

or

$$-\frac{1}{2\gamma_0}\frac{\partial\bar{\phi}}{\partial u} W_u(t)^{-1}$$

$$\times\left[\Lambda^J(t) + \Lambda^\psi(t)\gamma + \frac{\partial\bar{\phi}^T}{\partial u}\,\mu(t)\right] + c\bar{\phi} = 0$$

It is assumed that at points where $B(t)$ is not empty, $\partial\bar{\phi}/\partial u$ has full row rank, i.e., all constraint functions which are zero or positive are independent. Since $W_u(t)$ is nonsingular, the matrix

$$\Lambda^\phi(t) = \frac{\partial\bar{\phi}}{\partial u} W_u(t)^{-1}\frac{\partial\bar{\phi}^T}{\partial u} \tag{8-34}$$

is nonsingular. Therefore,

$$\mu(t) = -\Lambda^\phi(t)^{-1}$$

$$\times\left[\frac{\partial\bar{\phi}}{\partial u} W_u^{-1}\,(\Lambda^J + \Lambda^\psi\gamma) - 2\gamma_0 c\bar{\phi}\right]. \tag{8-35}$$

At points where $B(t)$ is empty, put $\partial\bar{\phi}/\partial u = \bar{\phi}(t) = 0$ and $\Lambda^\phi(t) = 1$. In this way, $\mu(t)$ is consistently defined by Eq. 8-35 for all t.

Substituting Eq. 8-35 into Eq. 8-32

$$\delta u(t) = -\frac{1}{2\gamma_0} W_u^{-1}\,(\Lambda^J + \Lambda^\psi\gamma)$$

$$+\frac{1}{2\gamma_0} W_u^{-1}\frac{\partial\bar{\phi}^T}{\partial u}\Lambda^{\phi^{-1}}$$

$$\times\frac{\partial\bar{\phi}}{\partial u} W_u^{-1}\,(\Lambda^J + \Lambda^\psi\gamma)$$

$$c W_u^{-1}\frac{\partial\bar{\phi}^T}{\partial u}\Lambda^{\phi^{-1}}\bar{\phi}$$

or

$$\delta u(t) = -\frac{1}{2\gamma_0} W_u^{-1} \times \left(I - \frac{\partial \bar{\phi}^T}{\partial u} \Lambda^{\phi^{-1}} \frac{\partial \bar{\phi}}{\partial u} W_u^{-1} \right) \times (\Lambda^J + \Lambda^\psi \gamma) - c W_u^{-1} \frac{\partial \bar{\phi}^T}{\partial u} \Lambda^{\phi^{-1}} \bar{\phi}. \tag{8-36}$$

where I is the identity matrix.

Substituting $\delta u(t)$ and δb from Eqs. 8-36 and 8-33 into Eq. 8-26 and then enforcing $\delta\bar{\psi} = -a\bar{\psi}$,

$$-\frac{1}{2\gamma_0} \ell^{\psi^T} W_b^{-1} (\ell^J + \ell^\psi \gamma) - c \int_{t^0}^{t^1} \Lambda^{\psi^T} \times W_u^{-1} \frac{\partial \bar{\phi}^T}{\partial u} \Lambda^{\phi^{-1}} \bar{\phi}\, dt - \frac{1}{2\gamma_0} \int_{t^0}^{t^1} \Lambda^{\psi^T} \times W_u^{-1} \left(I - \frac{\partial \bar{\phi}^T}{\partial u} \Lambda^{\phi^{-1}} \frac{\partial \bar{\phi}}{\partial u} W_u^{-1} \right) \times (\Lambda^J + \Lambda^\psi \gamma)\, dt = -a\bar{\psi}. \tag{8-37}$$

Defining

$$M_{\psi J} = \ell^{\psi^T} W_b^{-1} \ell^J + \int_{t^0}^{t^1} \Lambda^{\psi^T} W_u^{-1} \times \left(I - \frac{\partial \bar{\phi}^T}{\partial u} \Lambda^{\phi^{-1}} \frac{\partial \bar{\phi}}{\partial u} W_u^{-1} \right) \Lambda^J\, dt \tag{8-38}$$

$$M_{\psi\psi} = \ell^{\psi^T} W_b^{-1} \ell^\psi + \int_{t^0}^{t^1} \Lambda^{\psi^T} W_u^{-1} \times \left(I - \frac{\partial \bar{\phi}^T}{\partial u} \Lambda^{\phi^{-1}} \frac{\partial \bar{\phi}}{\partial u} W_u^{-1} \right) \Lambda^\psi\, dt \tag{8-39}$$

and

$$M_{\psi\phi} = \int_{t^0}^{t^1} \Lambda^{\psi^T} W_u^{-1} \frac{\partial \bar{\phi}^T}{\partial u} \Lambda^{\phi^{-1}} \bar{\phi}\, dt, \tag{8-40}$$

Eq. 8-37 becomes

$$\frac{1}{2\gamma_0} (M_{\psi J} + M_{\psi\psi} \gamma) + c M_{\psi\phi} = a\bar{\psi}. \tag{8-41}$$

Since $W_u(t)$ is positive definite so is W_u^{-1}, and there is a nonsingular matrix $s(t)$ such that $W_u^{-1}(t) = s^T(t)s(t)$. By direct multiplication, it may be verified that

$$y^T M_{\psi\psi} y = y^T \ell^{\psi^T} W_b^{-1} \ell^\psi y + \int_{t^0}^{t^1} y^T \Lambda^{\psi^T} W_u^{-1} \times \left(I - \frac{\partial \bar{\phi}^T}{\partial u} \Lambda^{\phi^{-1}} \frac{\partial \bar{\phi}}{\partial u} W_u^{-1} \right) \Lambda^\psi y\, dt$$

$$= y^T \ell^{\psi^T} W_b^{-1} \ell^\psi y + \int_{t^0}^{t^1} \left[\left(I - s \frac{\partial \bar{\phi}^T}{\partial u} \Lambda^{\phi^{-1}} \times \frac{\partial \bar{\phi}}{\partial u} s^T \right) s \Lambda^\psi y \right]^T \times \left[\left(I - s \frac{\partial \bar{\phi}^T}{\partial u} \Lambda^{\phi^{-1}} \frac{\partial \bar{\phi}}{\partial u} s^T \right) \times s \Lambda^\psi y \right] dt \geq 0. \tag{8-42}$$

Therefore, $M_{\psi\psi}$ is at least positive semidefinite. In the development that follows, it will be assumed that $M_{\psi\psi}$ is positive definite and, hence, nonsingular.

Eq. 8-41 is now solved for γ to obtain

$$\gamma = M_{\psi\psi}^{-1}\left[2\gamma_0(a\bar{\psi} - cM_{\psi\phi}) - M_{\psi J}\right]. \tag{8-43}$$

It should be noted that if the set of indices A is empty, $\bar{\psi}$ does not exist so $M_{\psi\psi}$ is not even defined. If, in this case, $M_{\psi\psi}$ is defined as one and $\bar{\psi}$ zero, then $\gamma = 0$ in Eqs. 8-41 and 8-36 reduces, appropriately. In this way, a single mathematical analysis holds in all cases.

Substituting Eq. 8-43 into Eqs. 8-33 and 8-36 yields

$$\delta u(t) = -W_u^{-1}\left(I - \frac{\partial\bar{\phi}^T}{\partial u}\Lambda^{\phi^{-1}}\frac{\partial\bar{\phi}}{\partial u}W_u^{-1}\right) \times \left[\frac{1}{2\gamma_0}\Lambda^J + \Lambda^\psi M_{\psi\psi}^{-1}(a\bar{\psi} - cM_{\psi\phi}) - \frac{1}{2\gamma_0}\Lambda^\psi M_{\psi\psi}^{-1}M_{\psi J}\right] - cW_u^{-1}\frac{\partial\bar{\phi}^T}{\partial u}\Lambda^{\psi^{-1}}\bar{\phi}$$

and

$$\delta b = -\frac{1}{2\gamma_0}W_b^{-1}\ell^J - W_b^{-1}\ell^\psi M_{\psi\psi}^{-1}(a\bar{\psi} - cM_{\psi\phi}) + \frac{1}{2\gamma_0}W_b^{-1}\ell^\psi M_{\psi\psi}^{-1}M_{\psi J}.$$

Defining

$$\delta u^1(t) = W_u^{-1}\left(I - \frac{\partial\bar{\phi}^T}{\partial u}\Lambda^{\phi^{-1}}\frac{\partial\bar{\phi}}{\partial u}W_u^{-1}\right) \times (\Lambda^J - \Lambda^\psi M_{\psi\psi}^{-1}M_{\psi J}) \tag{8-44}$$

$$\delta u^2(t) = -W_u^{-1}\left(I - \frac{\partial\bar{\phi}^T}{\partial u}\Lambda^{\phi^{-1}}\frac{\partial\bar{\phi}}{\partial u}W_u^{-1}\right) \times \Lambda^\psi M_{\psi\psi}^{-1}(a\bar{\psi} - cM_{\psi\phi}) - cW_u^{-1}\frac{\partial\bar{\phi}^T}{\partial u}\Lambda^{\phi^{-1}}\bar{\phi} \tag{8-45}$$

$$\delta b^1 = W_b^{-1}(\ell^J - \ell^\psi M_{\psi\psi}^{-1}M_{\psi J}) \tag{8-46}$$

and

$$\delta b^2 = -W_b^{-1}\ell^\psi M_{\psi\psi}^{-1}(a\bar{\psi} - cM_{\psi\phi}) \tag{8-47}$$

the expressions for $\delta u(t)$ and δb are simply

$$\delta u(t) = -\frac{1}{2\gamma_0}\delta u^1(t) + \delta u^2(t) \tag{8-48}$$

and

$$\delta b = -\frac{1}{2\gamma_0}\delta b^1 + \delta b^2. \tag{8-49}$$

The variations $\delta u(t)$ and δb from Eqs. 8-48 and 8-49 could now be substituted into Eq. 8-27 to determine γ_0. However, since dP has no real physical significance, one might just as well choose γ_0. It is interesting to note that the terms $\delta u^2(t)$ and δb^2 are not multiplied by an undetermined parameter. Further, note that each term in the definitions, Eqs. 8-45 and 8-47, of these quantities involves $\bar{\phi}$ and $\bar{\psi}$.

In fact, if $\bar{\phi}$ and $\bar{\psi}$ are zero or empty, $\delta u^2(t) = \delta b^2 = 0$. It appears that $\delta u^2(t)$ and δb^2 may be interpreted as making corrections in constraint errors, or keeping constraint functions from being violated. Actually, this and more is true.

Theorem 8-1: The following identities among $\delta u^1(t)$, $\delta u^2(t)$, δb^1, and δb^2 of Eqs. 8-44 through 8-47 hold:

1. $\delta b^{1^T} W_b \delta b^2 + \int_{t^0}^{t^1} \delta u^{1^T} W_u \delta u^2 dt = 0$

2. $\ell^{\psi^T} \delta b^2 + \int_{t^0}^{t^1} \Lambda^{\psi^T} \delta u^2 dt = -a\bar{\psi}$

3. $\ell^{\psi^T} \delta b^1 + \int_{t^0}^{t^1} \Lambda^{\psi^T} \delta u^1 dt = 0$

4. $\frac{\partial \bar{\phi}}{\partial u} \delta u^1 = 0$

5. $\frac{\partial \bar{\phi}}{\partial u} \delta u^2 = -c\bar{\phi}$

6. $-\ell^{J^T} \delta b^1 - \int_{t^0}^{t^1} \Lambda^{J^T} \delta u^1 dt < 0 .$

By considering the case when $\bar{\phi}$ and $\bar{\psi}$ are empty, it is clear that in order for δu, δb to be in the negative gradient direction of J (i.e., $\delta u = -\Lambda^J$ and $\delta b = -\ell^J$), $\gamma_0 > 0$ is required.

The six relationships of Theorem 8-1 give the designer an intuitive feel for the Method of Steepest Descent. First, Relation 1 states that the changes $\delta u^1(t)$, δb^1 and $\delta u^2(t)$, δb^2 are orthogonal. Relations 2 and 5 show that $\delta u^2(t)$, δb^2 provides the requested reduction in the constraint functions. Relations 3 and 4 show, as might be expected due to the orthogonality of Relation 1, that $\delta u^1(t)$, δb^1 has no effect on the constraint functions in $\bar{\phi}$ and $\bar{\psi}$. Finally, Relation 6, along with Eqs. 8-48 and 8-49, simply states that if $\bar{\psi} = 0$ and $\bar{\phi}(t) = 0$ then $\delta u(t)$, δb provides a reduction in J.

Before stating a computational algorithm, it is important to develop a test for convergence to the solution of the original problem. The procedure here will be to show, through use of the necessary conditions of Chapter 6, that as the solution of the original problem is approached, $\delta u^1(t)$ and δb^1 must approach zero.

By Theorem 6-7, at the solution of the problem, Eqs. 8-1 through 8-4, there are multipliers $\omega_i(t)$, $i = 1, \ldots, n$, ν_α, $\alpha = 1, \ldots, r$, and $\xi_\beta(t)$, $\beta = 1, \ldots q$ such that for

$$H = \omega^T f - f_0 - \nu^T L - \xi^T \phi \tag{8-50}$$

and

$$G = g_0 + \nu^T g \tag{8-51}$$

it is required that

$$\frac{d\omega}{dt} = -\frac{\partial H^T}{\partial x} \tag{8-52}$$

$$\frac{\partial H}{\partial u} = 0 \tag{8-53}$$

$$\frac{\partial G}{\partial b} \int_{t^0}^{t^1} \frac{\partial H}{\partial b} dt = 0 \tag{8-54}$$

$$\nu_\alpha \psi_\alpha = 0, \quad \alpha = r' + 1, \ldots, r \tag{8-55}$$

$$\xi_\beta(t)\phi_\beta(t) = 0, \ \beta = q' + 1, \ldots, q. \tag{8-56}$$

Corresponding to the definition of $\bar{\psi}$ and $\bar{\phi}$ in Eqs. 8-17 and 8-18, define $\bar{\nu}$, $\bar{L}$, $\bar{g}$, and $\bar{\xi}$ as containing only components of ν, L, g, and ξ, corresponding to elements of $\bar{\psi}$ and $\bar{\phi}$. In this notation – and due to Eqs. 8-55 and 8-56 – Eqs. 8-52, 8-53, and 8-54 become

$$\frac{d\omega^T}{dt} = -\omega^T \frac{\partial f}{\partial x} + \frac{\partial f_0}{\partial x} + \bar{\nu}^T \frac{\partial \bar{L}}{\partial x} \tag{8-57}$$

$$\omega^T \frac{\partial f}{\partial u} - \frac{\partial f_0}{\partial u} - \bar{\nu}^T \frac{\partial \bar{L}}{\partial u} - \bar{\xi}^T \frac{\partial \bar{\phi}}{\partial u} = 0 \tag{8-58}$$

$$\frac{\partial g_0}{\partial b} + \bar{\nu}^T \frac{\partial \bar{g}}{\partial b} - \int_{t^0}^{t^1} \left(\omega^T \frac{\partial f}{\partial b} - \frac{\partial f_0}{\partial b} - \bar{\nu}^T \frac{\partial \bar{L}}{\partial b} \right) dt = 0 . \tag{8-59}$$

Substituting from Eq. 8-11 into Eq. 8-57 yield

$$\frac{d\omega}{dt} + \frac{d\lambda^J}{dt} + \frac{d\lambda^\psi}{dt} \bar{\nu} = -\frac{\partial f^T}{\partial x} \omega - \frac{\partial f^T}{\partial x} \lambda^J - \frac{\partial f^T}{\partial x} \lambda^\psi \bar{\nu} \tag{8-60}$$

or

$$\frac{d}{dt}(\omega + \lambda^J + \lambda^\psi \bar{\nu}) = -\frac{\partial f^T}{\partial x}(\omega + \lambda^J + \lambda^\psi \bar{\nu}) . \tag{8-61}$$

Further conditions from Theorem 6-7 are, from Eq. 6-124,

$$\frac{\partial G^T}{\partial x^0} - \omega(t^0) = 0 \tag{8-62}$$

and

$$\frac{\partial G^T}{\partial x^1} + \omega(t^1) = 0 \tag{8-63}$$

Multiplying Eqs. 8-62 and 8-63 by δx^0 and δx^1 yields

$$\frac{\partial g_0}{\partial x^0} \delta x^0 + \bar{\nu}^T \frac{\partial \bar{g}}{\partial x^0} \delta x^0 - \omega^T(t^0)\delta x^0 = 0 \tag{8-64}$$

and

$$\frac{\partial g_0}{\partial x^1} \delta x^1 + \bar{\nu}^T \frac{\partial \bar{g}}{\partial x^1} \delta x^1 + \omega^T(t^1)\delta x^1 = 0 . \tag{8-65}$$

These equations hold for all δx^0 and δx^1. Adding Eqs. 8-64 and 8-65,

$$\frac{\partial g_0}{\partial x^0} \delta x^0 + \frac{\partial g_0}{\partial x^1} \delta x^1 + \bar{\nu}^T \left(\frac{\partial \bar{g}}{\partial x^0} \delta x^0 + \frac{\partial \bar{g}}{\partial x^1} \delta x^1 \right) - \omega^T(t^0)\delta x^0 + \omega^T(t^1)\delta x^1 = 0 . \tag{8-66}$$

Again, Eq. 8-66 holds for all δx^0 and δx^1.

Substituting terms from Eqs. 8-13 and 8-15 into Eq. 8-66,

$$\lambda^{J^T}(t^1)\delta x^1 - \lambda^{J^T}(t^0)\delta x^0 + \bar{\nu}^T \left[\lambda^{\psi^T}(t^1)\delta x^1 - \lambda^{\psi^T}(t^0)\delta x^0 \right] - \omega^T(t^0)\delta x^0 + \omega^T(t^1)\delta x^1 = 0 \tag{8-67}$$

for all δx^ψ and δx^1 satisfying the second equation of Eq. 8-8. By collecting terms, Eq. 8-67 becomes

$$\left[\omega^T(t^1) + \lambda^{J^T}(t^1) + \bar{\nu}^T \lambda^{\psi^T}(t^1)\right]\delta x^1 - \left[\omega^T(t^0) + \lambda^{J^T}(t^0) + \bar{\nu}\lambda^{\psi^T}(t^0)\right]\delta x^0 = 0 \tag{8-68}$$

for all δx^0 and δx^1 satisfying the second equation of Eq. 8-8.

Eqs. 8-61 and 8-68 constitute the boundary-value problem adjoint to Eq. 8-8, where the dependent variable is $(\omega + \lambda^J + \lambda^\psi \bar{\nu})$. Due to the assumed well posed nature of Eq. 8-2, the boundary-value problem of Eq. 8-8 has a unique solution for all $\delta u(t)$ and δb. It is shown in Ref. 2, Chapter 4, that in this case the adjoint boundary-value problem, Eqs. 8-61 and 8-68, has a unique null solution, i.e.,

$$\omega(t) + \lambda^J(t) + \lambda^\psi(t) = 0, \quad t^0 < t < t^1 . \tag{8-69}$$

Substituting for $\omega(t)$ from Eq. 8-69 into Eqs. 8-58 and 8-59 yields

$$-\left(\lambda^{J^T} + \bar{\nu}^T \lambda^{\psi^T}\right)\frac{\partial f}{\partial u} - \frac{\partial f_0}{\partial u} - \bar{\nu}^T \frac{\partial \bar{L}}{\partial u} - \bar{\xi}^T \frac{\partial \bar{\phi}}{\partial u} = 0 \tag{8-70}$$

and

$$\frac{\partial g_0}{\partial b} + \bar{\nu}^T \frac{\partial \bar{g}}{\partial b} - \int_{t^0}^{t^1}\left[-\left(\lambda^{J^T} + \bar{\nu}^T \lambda^{\psi^T}\right) \times \frac{\partial f}{\partial b} - \frac{\partial f_0}{\partial b} - \bar{\nu}^T \frac{\partial \bar{L}}{\partial b}\right] dt = 0 . \tag{8-71}$$

Premultiplying the transpose of Eq. 8-70 by $(\partial\bar{\phi}/\partial u)W_u^{-1}$ yields

$$\frac{\partial \bar{\phi}}{\partial u} W_u^{-1} \frac{\partial \bar{\phi}^T}{\partial u} \bar{\xi} = -\frac{\partial \bar{\phi}}{\partial u} W_u^{-1}\left(\frac{\partial f^T}{\partial u}\lambda^J + \frac{\partial f_0{}^T}{\partial u}\right) - \frac{\partial \bar{\phi}}{\partial u} W_u^{-1}\left(\frac{\partial f^T}{\partial u}\lambda^\psi + \frac{\partial \bar{L}^T}{\partial u}\right)\bar{\nu} . \tag{8-72}$$

The coefficient of $\bar{\xi}$ in Eq. 8-72 is just $\Lambda^\phi(t)$ of Eq. 8-34 which is nonsingular. Therefore,

$$\bar{\xi} = -\Lambda^{\phi^{-1}}\left[\frac{\partial \bar{\phi}}{\partial u} W_u^{-1}\left(\Lambda^J + \Lambda^\psi \bar{\nu}\right)\right] . \tag{8-73}$$

Substituting Eq. 8-73 into Eq. 8-71,

$$\frac{\partial g_0{}^T}{\partial b} + \frac{\partial g^T}{\partial b}\bar{\nu} + \int_{t^0}^{t^1}\left[\frac{\partial f^T}{\partial b}\left(\lambda^J + \lambda^{\psi^T}\bar{\nu}\right) + \frac{\partial f_0{}^T}{\partial b} + \frac{\partial \bar{L}^T}{\partial b}\bar{\nu}\right] dt = 0$$

or, in the notation of Eqs. 8-22 and 8-24,

$$\ell^J + \ell^\psi \bar{\nu} = 0 . \tag{8-74}$$

Substituting for $\bar{\xi}$ in Eq. 8-73 into Eq. 8-70,

$$-\Lambda^J - \Lambda^\psi \bar{\nu} + \frac{\partial \bar{\phi}^T}{\partial u}\Lambda^{\phi^{-1}}\frac{\partial \bar{\phi}}{\partial u} W_u^{-1}\Lambda^J + \frac{\partial \bar{\phi}^T}{\partial u}\Lambda^{\phi^{-1}}\frac{\partial \bar{\phi}}{\partial u} W_u^{-1}\Lambda^\psi \bar{\nu} = 0 . \tag{8-75}$$

Premultiplying Eq. 8-75 by $\Lambda^{\psi T} W_u^{-1}$ and integrating yields

$$\int_{t^0}^{t^1} \left(\Lambda^{\psi T} W_u^{-1} \Lambda^J - \Lambda^{\psi T} W_u^{-1} \times \frac{\partial \bar{\phi}^T}{\partial u} \Lambda^{\phi^{-1}} \frac{\partial \bar{\phi}}{\partial u} W_u^{-1} \Lambda^J \right) dt$$

$$+ \left[\int_{t^0}^{t^1} \left(\Lambda^{\psi T} W_u^{-1} \Lambda^{\psi} - \Lambda^{\psi T} W_u^{-1} \frac{\partial \bar{\phi}^T}{\partial u} \Lambda^{\phi^{-1}} \times \frac{\partial \bar{\phi}}{\partial u} W_u^{-1} \Lambda^{\psi} \right) dt \right] \bar{\nu} = 0 \quad (8\text{-}76)$$

Premultiplying Eq. 8-74 by $\ell^{\psi T} W_b^{-1}$ yields,

$$\ell^{\psi T} W_o^{-1} \ell^J + \ell^{\psi T} W_b^{-1} \ell^{\psi} \bar{\nu} = 0 . \quad (8\text{-}77)$$

Adding Eqs. 8-76 and 8-77 finally yields

$$M_{\psi J} + M_{\psi\psi} \bar{\nu} = 0$$

so

$$\bar{\nu} = - M_{\psi\psi}^{-1} M_{\psi J} . \quad (8\text{-}78)$$

Substituting $\bar{\nu}$ from Eq. 8-78 into Eq. 8-75,

$$\Lambda^J - \Lambda^{\psi} M_{\psi\psi}^{-1} M_{\psi J} - \frac{\partial \bar{\phi}^T}{\partial u} \Lambda^{\phi^{-1}} \frac{\partial \bar{\phi}}{\partial u} W_u^{-1} \Lambda^J + \frac{\partial \bar{\phi}^T}{\partial u} \Lambda^{\phi^{-1}} \frac{\partial \bar{\phi}}{\partial u} W_u^{-1} \Lambda^{\psi} M_{\psi\psi}^{-1} M_{\psi J} = 0$$

or,

$$\left(I - \frac{\partial \bar{\phi}^T}{\partial u} \Lambda^{\phi^{-1}} \frac{\partial \bar{\phi}}{\partial u} W_u^{-1} \right) \times \left(\Lambda^J - \Lambda^{\psi} M_{\psi\psi}^{-1} M_{\psi J} \right) = 0 . \quad (8\text{-}79)$$

Eq. 8-79 is the desired result, $\delta u^1(t) = 0$.

Substituting $\bar{\nu}$ from Eq. 8-78 into Eq. 8-74 yields

$$\ell^J - \ell^{\psi} M_{\psi\psi}^{-1} M_{\psi J} = 0$$

and this implies, by Eq. 8-46, that $\delta b^1 = 0$.

It is now possible to state a computational algorithm employing the results of the preceding analysis and discussion.

A Computational Algorithm:

Step 1. Make an engineering estimate $u^{(0)}(t)$, $b^{(0)}$, of the optimum design function and parameter.

Step 2. Solve Eq. 8-2 for $x^{(0)}$ corresponding to $u^{(0)}(t)$, $b^{(0)}$.

Step 3. Check constraints and form $\bar{\psi}$ and $\bar{\phi}(t)$ of Eqs. 8-17 and 8-18.

Step 4. Solve the differential equation 8-11 with h and the boundary conditions of Eqs. 8-13 and 8-15 to obtain $\lambda^J(t)$ and $\lambda^{\psi\alpha}(t)$, respectively.

Step 5. Compute $\Lambda^J(t)$, ℓ^J, $\Lambda^{\psi}(t)$, and ℓ^{ψ} in Eqs. 8-21 through 8-24 and $\Lambda^{\phi}(t)$ in Eq. 8-34.

Step 6. Choose the correction factors a and c in Eqs. 8-19 and 8-20.

Step 7. Compute $M_{\psi J}$, $M_{\psi\psi}$, and $M_{\psi\phi}$ in Eqs. 8-38, 8-39, and 8-40.

Step 8. Choose $\gamma_0 > 0$ and compute γ and $\mu(t)$ of Eqs. 8-43 and 8-35. If any components of γ with $\alpha > r'$, or $\mu(t)$ with $\beta > q'$, are negative, redefine $\tilde{\psi}$ and $\tilde{\phi}(t)$ by deleting corresponding terms and return to Step 5.

Step 9. Compute $\delta u^1(t)$, $\delta u^2(t)$, δb^1, and δb^2 of Eqs. 8-44 through 8-47.

Step 10. Compute

$$u^{(1)}(t) = u^{(0)}(t) - \frac{1}{2\gamma_0}\,\delta u^1(t) + \delta u^2(t)$$

$$b^{(1)} = b^{(0)} - \frac{1}{2\gamma_0}\,\delta b^1 + \delta b^2 .$$

Step 11. If the constraints are satisfied and $\delta u^1(t)$ and δb^1 are sufficiently small, terminate. Otherwise, return to Step 2 with $u^{(0)}$, $b^{(0)}$ being replaced by $u^{(1)}$, $b^{(1)}$.

An algorithm of this kind invariably involves a certain amount of computational art. The critical element of this algorithm is the choice of the parameter γ_0 in Step 8. Once the constraints are satisfied to acceptable accuracy, $\delta u^2(t)$ and δb^2 will be approximately zero and $1/(2\gamma_0)$ can be viewed as a step size in the direction $\delta u^1(t)$, δb^1. In this case the change in $u(t)$ and b is

$$\delta u(t) = -\frac{1}{2\gamma_0}\,\delta u^1(t)$$

$$\delta b = -\frac{1}{2\gamma_0}\,\delta b^1$$

Substituting these expressions into Eq. 8-25 and using Eqs. 8-44 and 8-46,

$$\begin{aligned}\delta J = -\frac{1}{2\gamma_0}\Bigg[\, & \ell^{J^T} W_b^{-1}\left(\ell^J - \ell^\psi M_{\psi\psi}^{-1} M_{\psi J}\right) \\ & + \int_{t^0}^{t^1} \Lambda^{J^T} W_u^{-1} \\ & \times \left(I - \frac{\partial\bar{\phi}}{\partial u}^T \Lambda^{\phi^{-1}} \frac{\partial\bar{\phi}}{\partial u} W_u^{-1}\right) \\ & \times \left(\Lambda^J - \Lambda^\psi M_{\psi\psi}^{-1} M_{\psi J}\right) dt\Bigg] \\ = \frac{1}{2\gamma_0} & \left(M_{JJ} - M_{\psi J}^T M_{\psi\psi}^{-1} M_{\psi J}\right)\end{aligned}$$

where

$$M_{JJ} = \ell^{J^T} W_b^{-1} \ell^J + \int_{t^0}^{t^1} \Lambda^{J^T} W_u^{-1} \times \left(I - \frac{\partial\bar{\phi}}{\partial u}^T \Lambda^{\phi^{-1}} \frac{\partial\bar{\phi}}{\partial u} W_u^{-1}\right) \Lambda^J\, dt.$$

With Eq. 8-80 it is possible to request a reasonable magnitude for δJ and compute the γ_0 which should give this reduction in J. In this way, it is possible to choose a reasonable γ_0. Experience with this method on structural design problems, of the kind discussed in the following chapter, has indicated that a request of 2% to 10% reduction in the cost function on the first iteration gives a value of γ_0 that yields convergence. Often, this value of γ_0 must be adjusted during the iterative process to prevent divergence or to speed convergence.

This matter of choosing step size in Step 8 requires a great deal more attention. With a little experience one can develop a "feel" for how to adjust γ_0 to get good convergence, even in complex problems. A feasible automatic method of choosing γ_0 is desirable for use on high-speed computers. No reliable method is known to the writer at this time.

8-3 A STEEPEST DESCENT METHOD FOR A GENERAL OPTIMAL DESIGN PROBLEM

8-3.1 THE PROBLEM CONSIDERED

The basic optimal design problem with fixed endpoints, no discontinuities, and no intermediate constraints was treated in the preceding paragraph. The problem considered here will be a generalization of that problem to include features such as variable endpoints, discontinuities, and intermediate constraints. The basic idea of the method of solution will be the same as in the preceding paragraph. Accounting for the additional features of this problem, however, introduces some complexity into the derivation of equations.

The problem to be treated here is to determine $u(t)$, $t^0 < t < t^\eta$, b, and $t^0, t^1, \ldots, t^\eta$ which minimize

$$J = g_0(b, t^j, x^j) + \int_{t^0}^{t^\eta} f_0[t, x(t)\, u(t), b]\, dt \tag{8-81}$$

subject to the conditions

$$\left.\begin{aligned} &\frac{dx}{dt} = f(t,x,u,b),\ t^0 < t < t^\eta,\ t \neq t^j \\ &\theta_s(t^0, x^0, t^\eta, x^\eta) = 0,\ s = 1, \ldots, n \end{aligned}\right\} \tag{8-82}$$

$$\Omega^i(t^j, x^j) = 0,\ i = 1, \ldots, \eta \tag{8-83}$$

$$\left.\begin{aligned} &\psi_\alpha = g_\alpha(b, t^j, x^j) + \int_{t^0}^{t^\eta} L_\alpha[t, x(t), u(t), b]\, dt = 0, \\ &\qquad \alpha = 1, \ldots, r' \\ &\psi_\alpha = g_\alpha(b, t^j, x^j) + \int_{t^0}^{t^\eta} L_\alpha[t, x(t), u(t), b]\, dt < 0, \\ &\qquad \alpha = r' + 1, \ldots, r \end{aligned}\right\} \tag{8-84}$$

and

$$\left.\begin{aligned} &\phi_\beta(t,u) = 0,\ t^0 < t < t^\eta, \\ &\qquad \beta = 1, \ldots, q' \\ &\phi_\beta(t,u) < 0,\ t^0 < t < t^\eta, \\ &\qquad \beta = q' + 1, \ldots, q. \end{aligned}\right\} \tag{8-85}$$

Note that this is just a special case of the problem of Def. 6-5. Equality constraints will be included in Eqs. 8-84 and 8-85 in a natural way during the development. It is assumed that for given $u(t), b, t^j$, and x^j the boundary-value problem, Eq. 8-82, has a continuous solution $x(t)$. If constraints of the form $\omega(t,x,u,b) < 0$ occur, they may be replaced by a constraint of the form

$$\int_{t^0}^{t^\eta} \{\omega[t, x(t), u(t), b] + |\omega[t, x(t), u(t), b]|\}\, dt = 0. \tag{8-86}$$

Constraints of the form of Eq. 8-85 are easily treated in a direct manner so they need not be reduced to the form of Eq. 8-86.

Just as in the method of par. 8-2, the idea here will be to estimate $u^{(0)}(t), b^{(0)}$, and $t^0, ..., t^\eta$, and then to allow small changes $\delta u(t)$ and δb. The object is to determine $\delta u(t)$ and δb which yield the greatest reduction in J and which satisfy the constraints.

8-3.2 THE EFFECT OF SMALL CHANGES IN DESIGN VARIABLES AND PARAMETERS

Before the optimum changes in $u(t)$ and b may be determined, the effect of these changes on J and ψ_α must be assessed. Since J and ψ_α have the same form, the expressions for change of a general functional

$$Q = g(b, t^j, x^j) + \int_{t^0}^{t^\eta} F[t, x(t), u(t), b]\, dt \tag{8-87}$$

will be determined and the result will be applied to J and ψ_α.

Expanding Q to first-order terms in the variables $u(.), b, t^j$, and x^j, yields

$$\begin{aligned} \delta Q = {} & \frac{\partial g}{\partial b}\delta b + \frac{\partial g}{\partial x^0}\Delta x^0 + ... + \frac{\partial g}{\partial x^j}\Delta x^j \\ & + ... + \frac{\partial g}{\partial x^\eta}\Delta x^\eta + \frac{\partial g}{\partial t^0}\delta t^0 + ... + \frac{\partial g}{\partial t^j}\delta t^j \\ & + ... + \frac{\partial g}{\partial t^\eta}\delta t^\eta - F(t^0 + 0)\delta t^0 + ... \\ & + [F(t^j - 0) - F(t^j + 0)]\,\delta t^j + ... \\ & + F(t^\eta - 0)\delta t^\eta \\ & + \int_{t^0}^{t^\eta}\left(\frac{\partial F}{\partial x}\delta x + \frac{\partial F}{\partial u}\delta u + \frac{\partial F}{\partial b}\delta b\right) dt, \end{aligned} \tag{8-88}$$

where Δx^j is the total change in x^j at the point t^j. Since $x(t)$ is to be continuous before and after the variation, the total change in $x(t)$ must be continuous at each point t. This requires that

$$\begin{aligned} \delta x^j(t^j - 0) + f(t^j - 0)\delta t^j &= \Delta x^j \\ &= \delta x^j(t^j + 0) + f(t^j + 0)\delta t^j \end{aligned} \tag{8-89}$$

$j = 0, 1, \ldots, \eta$, where $\delta x(t)$ and δt^j are indpendent changes in $x(t)$ and t^j.

The independent variation in $x(t)$, $\delta x(t)$, is related to $\delta u(t)$ and δb through the variational equation

$$\frac{d}{dt}(\delta x) = \frac{\partial f}{\partial x}\delta x + \frac{\partial f}{\partial u}\delta u + \frac{\partial f}{\partial b}\delta b, \qquad t^0 < t < t^\eta,\ t \neq t^j. \tag{8-90}$$

The boundary conditions, Eq. 8-82, require that

$$\frac{\partial \theta_s}{\partial x^0}\Delta x^0 + \frac{\partial \theta_s}{\partial t^0}\delta t^0 + \frac{\partial \theta_s}{\partial x^\eta}\Delta x^\eta + \frac{\partial \theta_s}{\partial t^\eta}\delta t^\eta = 0 \tag{8-91}$$

$s = 1, ..., n$. Finally, the relations of Eq. 8-83 require that

$$\frac{\partial \Omega^j}{\partial x^j}\Delta x^j + \frac{\partial \Omega^j}{\partial t^j}\delta t^j = 0, \qquad j = 0, 1, ..., \eta. \tag{8-92}$$

It is clear that the variations $\delta x(t)$, $\delta u(t)$, δb, δt^j, and Δx^j are not all independent.

In order to express δQ in terms of only

$\delta u(t)$ and δb which are to be determined, introduce the adjoint variable $\lambda(t)$ just as in Eq. 8-91,

$$\frac{d\lambda}{dt} = -\frac{\partial f^T}{\partial x}\lambda - \frac{\partial F^T}{\partial x}. \tag{8-93}$$

Integrating the identity below Eq. 8-91 from t^j to t^{j+1}, $0 < j, j+1 < \eta$, one obtains, just as in Eq. 8-92,

$$\lambda^T(t^{j+1}-0)\delta x(t^{j+1}-0)$$
$$-\lambda^T(t^j+0)\delta x(t^j+0)$$
$$= \int_{t^j}^{t^{j+1}} \left(-\frac{\partial F}{\partial x}\delta x + \lambda^T \frac{\partial f}{\partial u}\delta u\right.$$
$$\left. + \lambda^T \frac{\partial f}{\partial b}\delta b\right) dt. \tag{8-94}$$

Note that the boundary conditions on $\lambda(t)$ and the properties of $\lambda(t)$ at the points t^j, $j = 1, ..., \eta - 1$, have not yet been specified. These boundary and intermediate conditions on $\lambda(t)$ will be the major output of this subparagraph.

Summing all the formulas, Eq. 8-94, over $j = 0, 1, ..., \eta - 1$, one obtains

$$\int_{t^0}^{t^\eta} \frac{\partial F}{\partial x}\delta x\, dt = -\lambda^T(t^0+0)\delta x(t^0+0)$$
$$+ ... + \lambda^T(t^j - 0)\delta x(t^j - 0)$$
$$-\lambda^T(t_j+0)\delta x(t^j+0) + ... \tag{8-95}$$
$$+\lambda^T(t^\eta - 0)\delta x(t^\eta - 0)$$
$$+\int_{t^0}^{t^\eta}\left(\lambda^T\frac{\partial f}{\partial u}\delta u + \lambda^T\frac{\partial f}{\partial b}\delta b\right) dt$$

Or, using the definition of Δx^j, this is

$$\int_{t^0}^{t^\eta} \frac{\partial F}{\partial x}\delta x\, dt = -\lambda^T(t^0+0)\Delta x^0$$
$$+\lambda^T(t^0+0) f(t^0+0)\delta t^0$$
$$+ ... + [\lambda^T(t^j - 0)$$
$$-\lambda^T(t^j+0)]\Delta x^j$$
$$-[\lambda^T(t^j-0) f(t^j-0)$$
$$-\lambda^T(t^j+0) f(t^j+0)]\delta t^j$$
$$+ ... + \lambda^T(t^\eta - 0)\Delta x^\eta$$
$$-\lambda^T(t^\eta - 0) f(t^\eta - 0)\delta t^\eta$$
$$+\int_{t^0}^{t^\eta}\left(\lambda^T \frac{\partial f}{\partial u}\delta u\right.$$
$$\left.+\lambda^T\frac{\partial f}{\partial b}\delta b\right) dt.$$

Substituting from Eq. 8-95 into Eq. 8-88, yields

$$\delta Q = \frac{\partial g}{\partial b}\delta b + \left[\frac{\partial g}{\partial x^0} + \lambda^T(t^0)\right]\Delta x^0 + ...$$
$$+\left[\frac{\partial g}{\partial x^j} - \lambda^T(t^j - 0) + \lambda^T(t^j+0)\right]\Delta x^j$$
$$+ ... + \left[\frac{\partial g}{\partial x^\eta} - \lambda^T(t^\eta)\right]\Delta x^\eta$$
$$+\left[\frac{\partial g}{\partial t^0} - F(t^0+0)\right.$$
$$\left.\lambda^T(t^0+0) f(t^0+0)\right]\delta t^0$$
$$+ ... + \left[\frac{\partial g}{\partial t^j} + F(t^j - 0) - F(t^j+0)\right.$$

$$+\lambda^T(t^j-0)f(t^j-0)$$
$$\left.-\lambda^T(t^j+0)f(t^j+0)\right]\delta t^j$$
$$+\ldots+\left[\frac{\partial g}{\partial t^n}+F(t^n-0)\right.$$
$$\left.+\lambda^T(t^n-0)f(t^n-0)\right]\delta t^n$$
$$+\int_{t^0}^{t^n}\left[\left(\frac{\partial F}{\partial u}+\lambda^T\frac{\partial f}{\partial u}\right)\delta u\right.$$
$$\left.+\left(\frac{\partial F}{\partial b}+\lambda^T\frac{\partial f}{\partial b}\right)\delta b\right]dt\,. \tag{8-96}$$

The quantities Δx^j and δt^j appearing in Eq. 8-96 are required to satisfy the conditions of Eqs. 8-91 and 8-92. The objective now is to choose the boundary and intermediate conditions on $\lambda(t)$ so that Eq. 8-96 is independent of Δx^j and δt^j; i.e., so that

$$\left[\frac{\partial g}{\partial x^0}+\lambda^T(t^0)\right]\Delta x^0+\ldots$$
$$+\left[\frac{\partial g}{\partial x^j}-\lambda^T(t^j-0)+\lambda^T(t^j+0)\right]\Delta x^j$$
$$+\ldots+\left[\frac{\partial g}{\partial x^n}-\lambda^T(t^n)\right]\Delta x^n$$
$$+\left[\frac{\partial g}{\partial t^0}-F(t^0+0)\right.$$
$$\left.-\lambda^T(t^0+0)f(t^0+0)\right]\delta t^0$$
$$+\ldots+\left[\frac{\partial g}{\partial t^j}+F(t^j-0)\right.$$
$$-F(t^j+0)+\lambda^T(t^j-0)f(t^j-0)$$
$$\left.-\lambda^T(t^j+0)f(t^j+0)\right]\delta t^j$$
$$+\ldots+\left[\frac{\partial g}{\partial t^n}+F(t^n-0)\right.$$
$$\left.+\lambda^T(t^n-0)f(t^n-0)\right]\delta t^n=0 \tag{8-97}$$

for all Δx^j and δt^j satisfying Eqs. 8-90 and 8-92.

In order to determine conditions on $\lambda(t^j)$ based on Eq. 8-97, a Lemma is required.

Lemma 8-1: For A, B_i, $i = 1, \ldots, m < n$ in R^n if $A^T x = 0$ for all x in R^n such that

$$B_i^T x = 0,\ i = 1, \ldots, m$$

then there exist constants ω_i such that

$$A^T x+\sum_{i=1}^{m}\omega_i B_i^T x=0 \tag{8-98}$$

for all x in R^n.

For proof of this Lemma see Ref. 3, page 12.

Consider the expression of Eq. 8-97 as $A^T x$, where the components of A depend on the values of $\lambda(t^j)$ and $x = \left(\Delta x^{0T}, \ldots, \Delta x^{nT}, \delta t^0, \ldots, \delta t^n\right)^T$. The equalities $B_i^T x = 0$ are just Eqs. 8-91 and 8-92. Denoting the multipliers ω_i, as $\tau_s, s = 1, \ldots, n$, and $\gamma_j, j = 0, 1, \ldots, \eta$, Eq. 8-98 becomes

$$\left[\frac{\partial g}{\partial x^0}+\lambda^T(t^0+0)+\sum_{s=1}^{n}\tau_s\frac{\partial\theta_s}{\partial x^0}\right.$$
$$\left.+\gamma_0\frac{\partial\Omega^0}{\partial x^0}\right]\Delta x^0$$
$$+\ldots+\left[\frac{\partial g}{\partial x^j}-\lambda^T(t^j-0)+\lambda^T(t^j+0)\right.$$

$$+ \gamma_j \frac{\partial \Omega^j}{\partial x^j} \Big] \Delta x^j$$

$$+ \ldots + \Big[\frac{\partial g}{\partial x^\eta} - \lambda^T (t^\eta - 0)$$

$$+ \sum_{s=1}^{n} \tau_s \frac{\partial \theta_s}{\partial x^\eta} + \gamma_\eta \frac{\partial \Omega^\eta}{\partial x^\eta} \Big] \Delta x^\eta$$

$$+ \Big[\frac{\partial g}{\partial t^0} - F(t^0 + 0) - \lambda^T (t^0 + 0) f(t^0 + 0)$$

$$+ \sum_{s=1}^{n} \tau_s \frac{\partial \theta_s}{\partial t^0} + \gamma_0 \frac{\partial \Omega^0}{\partial t^0} \Big] \delta t^0$$

$$+ \ldots + \Big[\frac{\partial g}{\partial t^j} + F(t^j - 0) - F(t^j - 0)$$

$$+ \lambda^T (t^j - 0) f(t^j - 0)$$

$$- \lambda^T (t^j + 0) f(t^j + 0) + \gamma_j \frac{\partial \Omega^j}{\partial t^j} \Big] \delta t^j$$

$$+ \ldots + \Big[\frac{\partial g}{\partial t^\eta} + F(t^\eta - 0)$$

$$+ \lambda^T (t^\eta - 0) f(t^\eta - 0) + \sum_{s=1}^{n} \tau_s \frac{\partial \theta_s}{\partial t^\eta}$$

$$+ \gamma_\eta \frac{\partial \Omega^\eta}{\partial t^\eta} \Big] \delta t^\eta = 0 \tag{8-99}$$

for all $\Delta x^j, \delta t^j, j = 0, 1, \ldots, \eta$. Therefore,

$$\frac{\partial g}{\partial x^0} + \lambda^T (t^0 + 0) + \sum_{s=1}^{n} \tau_s \frac{\partial \theta_s}{\partial x^0}$$

$$+ \gamma_0 \frac{\partial \Omega^0}{\partial x^0} = 0 \tag{8-100}$$

$\vdots$

$$\frac{\partial g}{\partial x^j} - \lambda^T (t^j - 0) + \lambda^T (t^j + 0) + \gamma_j \frac{\partial \Omega^j}{\partial x^j} = 0 \tag{8-101}$$

$\vdots$

$$\frac{\partial g}{\partial x^\eta} - \lambda^T (t^\eta - 0) + \sum_{s=1}^{n} \tau_s \frac{\partial \theta_s}{\partial x^\eta}$$

$$+ \gamma_\eta \frac{\partial \Omega^\eta}{\partial x^\eta} = 0 \tag{8-102}$$

$$\frac{\partial g}{\partial t^0} - F(t^0 + 0) - \lambda^T (t^0 + 0) f(t^0 + 0)$$

$$+ \sum_{s=1}^{n} \tau_s \frac{\partial \theta_s}{\partial t^0} + \gamma_0 \frac{\partial \Omega^0}{\partial t^0} = 0 \tag{8-103}$$

$\vdots$

$$\frac{\partial g}{\partial t^j} + F(t^j - 0) - F(t^j - 0)$$

$$+ \lambda^T (t^j - 0) f(t^j - 0)$$

$$- \lambda^T (t^j + 0) f(t^j + 0)$$

$$+ \gamma_j \frac{\partial \Omega_j}{\partial t^j} = 0 \tag{8-104}$$

$\vdots$

$$\frac{\partial g}{\partial t^\eta} + F(t^\eta - 0) + \lambda^T (t^\eta - 0) f(t^\eta - 0)$$

$$+ \sum_{s=1}^{n} \tau_s \frac{\partial \theta_s}{\partial t^\eta} + \gamma_\eta \frac{\partial \Omega^\eta}{\partial t^\eta} = 0 \,. \tag{8-105}$$

The object is now to eliminate the γ_i and τ_s in order to obtain explicit conditions on $\lambda(t)$. Postmultiplying Eq. 8-100 by $f(t^0 + 0)$ and adding Eq. 8-103 yields

$$\dot{g}(t^0 + 0) - F(t^0 + 0) + \sum_{s=1}^{n} \tau_s \dot{\theta}_s (t^0 + 0)$$

$$+ \gamma_0 \dot{\Omega}^0 (t^0 + 0) = 0 \tag{8-106}$$

where

$$\dot{g}(t^i \pm 0) = \frac{\partial g}{\partial t^i}$$

$$+ \frac{\partial g}{\partial x^i} f[t^i \pm 0, x(t^i \pm 0), u(t^i \pm 0), b] \tag{8-107}$$

$$\dot{\theta}_s(t^l \pm 0) = \frac{\partial \theta_s}{\partial t^l} + \frac{\partial \theta_s}{\partial x^l} f[t^l \pm 0, x(t^l \pm 0), u(t^l \pm 0), b] \tag{8-108}$$

and

$$\dot{\Omega}^l(t^l \pm 0) = \frac{\partial \Omega^l}{\partial t^l} + \frac{\partial \Omega^l}{\partial x^l} f[t^l \pm 0, x(t^l \pm 0), u(t^l \pm 0), b] \,. \tag{8-109}$$

Since $\Omega^0(t^0, x^0) = 0$ is to determine t^0, the total derivat[illegible] with respect to t^0, $\dot{\Omega}^0(t^0+0)$, should not be zero. Therefore, γ_0 may be determined as

$$\gamma_0 = -\frac{1}{\dot{\Omega}^0(t^0+0)} \times \left[\dot{g}(t^0+0) - F(t^0+0) + \sum_{s=1}^{n} \tau_s \dot{\theta}_s(t^0+0)\right] . \tag{8-110}$$

Substituting Eq. 8-110 into Eq. 8-100 yields

$$\lambda(t^0+0) = -\frac{\partial g^T}{\partial x^0} - \sum_{s=1}^{n} \tau_s \frac{\partial \theta_s^{0\,T}}{\partial x^0} + \frac{1}{\dot{\Omega}^0(t^0+0)}\left[\dot{g}(t^0+0) - F(t^0+0) + \sum_{s=1}^{n} \tau_s \dot{\theta}_s(t^0+0)\right]\frac{\partial \Omega^{0\,T}}{\partial x^0} \tag{8-111}$$

This is then a boundary condition on $\lambda(t)$ at t^0.

In exactly the same way, postmultiplying Eq. 8-102 by $f(t^n - 0)$ and adding to Eq. 8-105 yields

$$\dot{g}(t^n - 0) + F(t^n - 0) + \sum_{s=1}^{n} \tau_s \dot{\theta}_s(t^n - 0) + \gamma_\eta \dot{\Omega}^n(t^n - 0) = 0 \,.$$

Solving for γ_η and substituting into Eq. 8-10[illegible] yields

$$\lambda(t^n - 0) = \frac{\partial g^T}{\partial x^n} + \sum_{s=1}^{n} \tau_s \frac{\partial \theta_s^T}{\partial x^n} - \frac{1}{\dot{\Omega}^n(t^n - 0)}\left[\dot{g}(t^n - 0) + F(t^n - 0) + \sum_{s=1}^{n} \tau_s \dot{\theta}_s(t^n - 0)\right]\frac{\partial \Omega^{n\,T}}{\partial x^n} \tag{8-112}$$

where $\dot{g}(t^n - 0)$ and $\dot{\theta}_s(t^n - 0)$ are defined in Eqs. 8-107 and 8-108. Eq. 8-112 then serves as a boundary condition on $\lambda(t)$ at t^n.

In order to use Eqs. 8-111 and 8-112 as explicit conditions on $\lambda(t)$ at t^0 and t^n, the parameters τ_s must be eliminated. This may be accomplished by alegbraic manipulation in particular cases. To illustrate this idea on a problem which has been treated extensively in the literature (Refs. 5,7,8,9), consider the case in which a full set of initial conditions is given, i.e.,

$$\theta_s(t^0, x^0) = 0, \qquad s = 1, \ldots, n \,. \tag{8-113}$$

In this case, Eq. 8-112 yields

$$\lambda(t^n - 0) = \frac{\partial g^T}{\partial x^n} - \frac{1}{\dot{\Omega}^n(t^n - 0)} \times \left[\dot{g}(t^n - 0) + F(t^n - 0)\right] \frac{\partial \Omega^{n\,T}}{\partial x^n} \,. \tag{8-114}$$

Eq. 8-111, on the other hand, is just a vector equation with n components which determines $\tau_1, ..., \tau_n$. It gives no explicit information on $\lambda(t^0 + 0)$.

Now the boundary conditions on $\lambda(t)$ have been determined. It remains, however, to determine jump conditions on $\lambda(t)$ at the intermediate points t^j. Postmultiplying Eq. 8-101 by $f(t^j \pm 0)$ and adding to Eq. 8-104 yields

$$
\begin{aligned}
&\dot{g}(t^j \pm 0) + F(t^j - 0) - F(t^j + 0) \\
&\quad - \lambda^T(t^j - 0)\,[f(t^j \pm 0) - f(t^j - 0)] \\
&\quad + \lambda^T(t^j + 0)\,[f(t^j \pm 0) - f(t^j + 0)] \\
&\quad + \gamma_j \dot{\Omega}^j(t^j \pm 0) = 0 \qquad (8\text{-}115)
\end{aligned}
$$

where the notation of Eqs. 8-107 through 8-109 has been used. The choice of limit from the right or left (the plus or minus sign, respectively) in Eq. 8-115 is left open for now. One or the other alternative will be chosen for computational convenience.

It is assumed that the condition $\Omega^j(t^j, x^j) = 0$ determines t^j as a function of x^j, so it is required that the total derivative of Ω^j with respect to t^j, $\dot{\Omega}^j(t^j \pm 0)$, not be zero. Therefore, from Eq. 8-115,

$$
\begin{aligned}
\gamma_j = -\frac{1}{\dot{\Omega}^j(t^j \pm 0)} \Big\{ &\dot{g}(t^j \pm 0) + F(t^j - 0) \\
&- F(t^j + 0) \\
&- \lambda^T(t^j - 0)\,[f(t^j \pm 0) - f(t^j - 0)] \\
&+ \lambda^T(t^j + 0)\,[f(t^j \pm 0) - f(t^j + 0)] \Big\} .
\end{aligned}
$$

Substituting this expression into Eq. 8-101,

$$
\begin{aligned}
\lambda(t^j + 0) - \lambda(t^j - 0) = &-\frac{\partial g^T}{\partial x^j} \qquad (8\text{-}116) \\
&+ \frac{1}{\dot{\Omega}^j(t^j \pm 0)} \Big\{ \dot{g}(t^j \pm 0) + F(t^j - 0) \\
&- F(t^j + 0) \\
&- \lambda^T(t^j - 0)\,[f(t^j \pm 0) - f(t^j - 0)] \\
&+ \lambda^T(t^j + 0)\,[f(t^j \pm 0) - f(t^j + 0)] \Big\} \frac{\partial \Omega^{j^T}}{\partial x^j} .
\end{aligned}
$$

This equation, the boundary conditions at Eqs. 8-111 and 8-114, and the differential equation, Eq. 8-93, are to determine the adjoint variable, $\lambda(t)$, $t^0 \leq t \leq t^n$. The boundary and intermediate conditions on $\lambda(t)$ were constructed so that Eq. 8-97 holds and in turn, Eq. 8-96 becomes

$$
\delta Q = \frac{\partial g}{\partial b}\delta b + \int_{t^0}^{t^n} \left[\left(\frac{\partial F}{\partial u} + \lambda^T \frac{\partial f}{\partial u} \right) \delta u(t) + \left(\frac{\partial F}{\partial b} + \lambda^T \frac{\partial f}{\partial b} \right) \delta b \right] dt
$$

or

$$
\delta Q = \left[\frac{\partial g}{\partial b} + \int_{t^0}^{t^n} \left(\frac{\partial F}{\partial b} + \lambda^T \frac{\partial f}{\partial b} \right) dt \right] \delta b + \int_{t^0}^{t^n} \left(\frac{\partial F}{\partial u} + \lambda^T \frac{\partial f}{\partial u} \right) \delta u(t)\, dt. \qquad (8\text{-}117)
$$

This equation meets the objective of this subparagraph, namely, determination of the dependence of δQ on δb and $\delta u(t)$ explicitly. Since Q was any functional, this result can be applied to the particular functionals of the present problem, J and ψ_α. To obtain δJ and

$\delta\psi_\alpha$, define $\lambda^J(t)$ and $\lambda^{\psi_\alpha}(t)$ as the solution of Eqs. 8-93, 8-111, 8-114, and 8-116 with

$$g = g_0 \text{ and } F = f_0 \text{ for } \lambda^J(t) \quad (8\text{-}118)$$

and

$$g = g_\alpha \text{ and } F = L_\alpha \text{ for } \lambda^{\psi_\alpha}(t). \quad (8\text{-}119)$$

In this notation, Eq. 8-117 yields

$$\delta J = \left\{ \frac{\partial g_0}{\partial b} + \int_{t^0}^{t^n} \left[\frac{\partial f_0}{\partial b} + \lambda^{J^T}(t) \frac{\partial f}{\partial b} \right] dt \right\} \delta b + \int_{t^0}^{t^n} \left[\frac{\partial f_0}{\partial u} + \lambda^{J^T}(t) \frac{\partial f}{\partial u} \right] \delta u(t) dt$$

and

$$\delta\psi_\alpha = \left\{ \frac{\partial g_\alpha}{\partial b} + \int_{t^0}^{t^n} \left[\frac{\partial L_\alpha}{\partial b} + \lambda^{\psi_\alpha^T}(t) \frac{\partial f}{\partial b} \right] dt \right\} \delta b + \int_{t^0}^{t^n} \left[\frac{\partial L_\alpha}{\partial u} + \lambda^{\psi_\alpha^T}(t) \frac{\partial f}{\partial u} \right] \delta u(t) dt .$$

For a more compact notation define

$$\ell^J = \frac{\partial g_0^T}{\partial b} + \int_{t^0}^{t^n} \left[\frac{\partial f_0^T}{\partial b} + \frac{\partial f^T}{\partial b} \lambda^J(t) \right] dt \quad (8\text{-}120)$$

$$\Lambda^J(t) = \frac{\partial f_0^T}{\partial u} + \frac{\partial f^T}{\partial u} \lambda^J(t) \quad (8\text{-}121)$$

$$\ell^{\psi_\alpha} = \frac{\partial g_\alpha^T}{\partial b} + \int_{t^0}^{t^n} \left[\frac{\partial L_\alpha^T}{\partial b} + \frac{\partial f^T}{\partial b} \lambda^{\psi_\alpha}(t) \right] dt \quad (8\text{-}122)$$

and

$$\Lambda^{\psi_\alpha}(t) = \frac{\partial L_\alpha^T}{\partial u} + \frac{\partial f^T}{\partial u} \lambda^{\psi_\alpha}(t). \quad (8\text{-}123)$$

In this notation

$$\delta J = \ell^{J^T} \delta b + \int_{t^0}^{t^n} \Lambda^{J^T}(t) \delta u(t) dt \quad (8\text{-}124)$$

and

$$\delta\psi_\alpha = \ell^{\psi_\alpha^T} \delta b + \int_{t^0}^{t^n} \Lambda^{\psi_\alpha^T}(t) \delta u(t) dt . \quad (8\text{-}125)$$

The problem of this paragraph is now in approximately the same state as the problem of par. 8-2 was in Eqs. 8-14 and 8-16. Before proceeding to derive a steepest descent algorithm, however, several comments are in order.

First, the choice of limit from the right or left was not made in Eq. 8-116. This choice is generally made depending on the distribution of boundary conditions on the state variable. If most of the boundary conditions on $x(t)$ are given at t^0, for example, then most of the boundary conditions on $\lambda(t)$ will be given at t^n. Since the adjoint equations are linear, superposition techniques may be used to solve the boundary-value problem. These techniques involve several integrations of Eq. 8-93 from t^n to t^0 with different starting conditions at t^n. These integrations must account for the jump condition, Eq 8-116. The integration then proceeds from the right and Eq. 8-116 should then be used to determine $\lambda(t^j - 0)$ in terms of $\lambda(t^j + 0)$ so that the integration may continue. For this reason, the minus sign is chosen in Eq. 8-116 so that

$$\lambda(t^j - 0) = \lambda(t^j + 0) + \frac{\partial g^T}{\partial x^j}$$

$$- \frac{1}{\dot{\Omega}^j(t^j - 0)} \dot{g}(t^j - 0)$$

$$+ F(t^j - 0) - F(t^j + 0)$$

$$+ \lambda(t^j + 0)[f(t^j - 0)$$

$$- f(t^j + 0)] \frac{\partial \Omega^{j\,T}}{\partial x^j} \quad . \quad (8\text{-}126)$$

Since the state equations have previously been integrated, $\dot{g}(t^j - 0)$ and $\dot{\Omega}^j(t^j - 0)$ can be computed in Eqs. 8-107 and 8-109.

The second matter that requires discussion is the determination of t^j and its variation, $j = 0, 1, ..., \eta$. If the state equations form an initial-value problem (all initial conditions given) then one can make an estimate for $u(t)$ and b and integrate Eq. 8-93 from t^0 toward t^η (or t^η toward t^0 if all boundary conditions are given at t^η). As the integration progresses, $\Omega^j(t,x)$ can be monitored and the value of t for which it is zero is called t^j. The situation is not so easy in case the state equations form a boundary-value problem.

One method of determining t^j requires that a reasonable estimate of t^j be available, perhaps from engineering intuition or preliminary analysis. The state equations are then integrated using the engineering estimates for $u(t)$ and b. It is likely that for the solution $x(t)$,

$$\Omega^j[t^j, x(t^j)] \neq 0$$

One might argue that the function $x(t)$ is close to the actual state and examine the effect on $\Omega^j[t^j,x(t^j)]$ of altering t^j, i.e.,

$$\delta\Omega^j[t^j,x(t^j)] = \frac{\partial \Omega^j}{\partial t^j}\,\delta t^j$$

$$+ \frac{\partial \Omega^j}{\partial x}\,\frac{dx}{dt}\,\delta t^j = \dot{\Omega}^j(t^j \pm 0)\,\delta t^j \quad (8\text{-}127)$$

where the plus or minus sign is chosen depending on whether δt^j should be positive or negative to make $\Omega^j(t^j + t^j), x(t^j + t^j) = 0$. The change δt^j is then chosen and if it is not too large, the state equations need not be re-integrated. This argument corresponds to a Newton-type algorithm for the determination of t^j. This procedure should be used after every variation in $u(t)$ and b and subsequent integration of the state equations, since $x(t)$ will be altered with an accompanying alteration in t^j.

8-3.3 A STEEPEST DESCENT COMPUTATIONAL ALGORITHM

The problem of determining $\delta u(t)$ and δb which reduce J and satisfy other constraints will now be solved just as in par. . As in the preceding paragraph, if some $\phi_\beta[u(t)]$ or ψ_α is less than zero, it will be ignored. If, on the other hand, $\psi_\alpha > 0$ or $\phi_\beta[u(t)] > 0$, then it will be required that

$$\delta\psi_\alpha = -a\psi_\alpha$$

and

$$\delta\phi_\beta = c\phi_\beta$$

where $0 < a < 1$ and $0 < c < 1$

Just as in par. 8-2, define two sets of indices

$$A = \left\{ \alpha \mid \psi_\alpha [x^{(0)}, u^{(0)}, b^{(0)}] > 0 \right\}$$

and

$$B(t) = \left\{ \beta \mid \phi_\beta [t, u^{(0)}(t)] > 0 \right\}$$

where $u^{(0)}(t)$ and $b^{(0)}$ are the beginning estimate of the design variable and design parameter, respectively, and $x^{(0)}(t)$ is the associated solution of the state equations. Further, define the column vectors of constraint functions

$$\bar{\psi} = \begin{bmatrix} \psi_\alpha \\ \\ \alpha \in A \end{bmatrix} \quad (8\text{-}128)$$

and

$$\bar{\phi}(t) = \begin{bmatrix} \phi_\beta [t, u^{(0)}(t)] \\ \\ \beta \in B \end{bmatrix} . \quad (8\text{-}129)$$

By the argument of par. 8-2, it will be required that

$$\delta \bar{\psi} = - a \bar{\psi} \quad (8\text{-}130)$$

and

$$\delta \bar{\phi}(t) = - c \bar{\phi}(t) . \quad (8\text{-}131)$$

Using the notation of Eqs. 8-122 and 8-123, define the matrices

$$\ell^{\psi} = \begin{bmatrix} \ell^{\psi_\alpha} \\ \\ \alpha \in A \end{bmatrix} \quad (8\text{-}132)$$

and

$$\Lambda^{\psi}(t) = \begin{bmatrix} \Lambda^{\psi_\alpha}(t) \\ \\ \alpha \in A \end{bmatrix} . \quad (8\text{-}133)$$

That is, the columns of ℓ^ψ and $\Lambda^\psi(t)$ are ℓ^{ψ_α} and $\Lambda^{\psi_\alpha}(t)$ for those α with $\psi_\alpha > 0$. Now,

$$\delta J = \ell^{J^T} \delta b + \int_{t^0}^{t^n} \Lambda^{J^T}(t) \delta u(t) dt \quad (8\text{-}134)$$

and

$$\delta \bar{\psi} = \ell^{\psi^T} \delta b + \int_{t^0}^{t^n} \Lambda^{\psi^T}(t) \delta u(t) dt . \quad (8\text{-}135)$$

The problem of this paragraph is now to find $\delta u(t)$ and δb to minimize δJ, subject to Eqs. 8-130 and 8-131. Although the symbols have a slightly different origin, this problem is precisely the same as that given by Eqs. 8-19, 8-20, and 8-25 of par. 8-2. All the analysis required to determine $\delta u(t)$ and δb follows and Theorem 8-1 holds. The only difference is that t^1 in Theorem 8-1 must be interpreted as t^n in the present problem.

The algorithm of par. 8-2 may now be given with references to equations of this paragraph.

Algorithm.

Step 1. Make an engineering estimate $u^{(0)}(t)$, $b^{(0)}$ of the optimum design function and parameter.

Step 2. Estimate t^0, t^1 ..., t^n, and solve Eq. 8-82 for $x^0(t)$

Step 3. Adjust t^j as required by the discussion below Eq. 8-127 and recompute $x^0(t)$ if required

Step 4. Check constraints and form $\tilde{\psi}$ and $\tilde{\phi}(t)$ of Eqs. 8-128 and 8-129.

Step 5. Solve the differential equation, Eq. 8-93, with boundary and intermediate conditions of Eqs. 8-111, 8-114, and 8-116. The solutions corresponding to the functions, Eqs. 8-118 and 8-119, yield $\lambda^J(t)$ and $\lambda^{\psi\alpha}(t)$, respectively.

Step 6. Compute ℓ^J, $\Lambda^J(t)$, ℓ^ψ, $\Lambda^\psi(t)$, and $\Lambda^\phi(t)$, in Eqs. 8-120, 8-121, 8-132, 8-133, and 8-34, respectively.

Step 7. Choose the correction factors a and c in Eqs. 8-130 and 8-131.

Step 8. Compute $M_{\psi J}$, $M_{\psi\psi}$, and $M_{\psi\phi}$ in Eqs. 8-38, 8-39, and 8-40.

Step 9. Choose $\gamma_0 > 0$ and compute γ and $\mu(t)$ of Eqs. 8-43 and 8-35. If any components of γ, with $\alpha > r'$, or $\mu(t)$, with $\beta > q'$ are negative, redefine $\tilde{\psi}$ and $\tilde{\phi}(t)$ by deleting corresponding terms and return to Step 6.

Step. 10. Compute $\delta u^1(t)$, $\delta u^2(t)$, δb^1, and δb^2 of Eqs. 8-44 through 8-47

Step. 11. Compute

$$u^{(1)}(t) = u^{(0)}(t) - \frac{1}{2\gamma_0}\,\delta u^1(t) + \delta u^2(t)$$

$$b^{(1)} = b^{(0)} - \frac{1}{2\gamma_0}\,\delta b^1 + \delta b^2 \,.$$

Step 12. If the constraints are satisfied and $\delta u^1(t)$ and δb^1 are sufficiently small, terminate. Otherwise, proceed to Step 13.

Step 13. Adjust $t^0, t^1, ..., t^n$ as required by the discussion below Eq. 8-127. Return to Step 2 with $u^{(0)}$, $b^{(0)}$ being replaced by $u^{(1)}$ and $b^{(1)}$.

For an alternate development of the algorithm in the special case of a full set of initial conditions, see Refs. 5 and 7. Several example problems are solved in Ref. 5.

8-4 STEEPEST DESCENT PROGRAMMING FOR A CLASS OF SYSTEMS DESCRIBED BY PARTIAL DIFFERENTIAL EQUATIONS

8-4.1 THE CLASS OF PROBLEMS CONSIDERED

Thus far, all problems considered have had their state variable specified by algebraic equations or boundary-value problems in ordinary differential equations. It is possible, however, that the state of the system being considered is governed by a boundary-value problem with partial differential equations. In such cases, the state and design variables are functions of more than one independent variable. One may then think of the design variable as being distributed over an area, volume, or higher dimensional space For this reason, such problems have been described as distributed parameter systems

A great deal of work has been done on distributed parameter systems which have a time-like variable (Refs. 12,13); i.e., a variable which makes the governing differential equation hyperbolic or parabolic. In this paragraph, consideration will be limited to static problems such as equilibrium of plates, shells, etc. These problems are described by linear, elliptic, partial differential equations (Ref. 1).

The boundary-value problem treated will be denoted

$$L(u,b)[z] = Q(x,u,b),\ x\in\Omega \quad (8\text{-}136)$$

$$B(v,b)[z] = q(x,v,b),\ x\in\Gamma \quad (8\text{-}137)$$

where $x = (x_1, x_2, ..., x_k)^T$ is the independent variable which ranges over the domain Ω in R^k with boundary Γ. The vector $u(x) = [u_1(x), ..., u_m(x)]^T$ is the design variable over Ω, $v(x) = [v_1(x), ..., v_r(x)]^T$ is the design variable over Γ (boundary design), and $b = (b_1, ..., b_\nu)^T$ is the design parameter. The state variable $z(x) = [z_1(x), ..., z_n(x)]^T$ is to be determined by the boundary-value problem, Eqs. 8-136 and 8-137, which is linear once u, v, and b are specified. It is important to note, however, that the problem depends in a nonlinear way on u, v, and b.

An example of the form of the differential operators $L(u,b)[z]$ and $B(v,b)[z]$ is

$$L(u,b)[z] = \sum_{|\alpha|<\eta} a_\alpha(x,u,b) \times \frac{\partial^{|\alpha|} z}{\partial x_1^{\alpha_1} \ldots \partial x_k^{\alpha_k}},\ x\in\Omega \quad (8\text{-}138)$$

and

$$B(v,b)[z] = \sum_{|\beta|<\eta} b_\beta(x,v,b) \times \frac{\partial^{|\beta|} z}{\partial x_1^{\beta_1} \ldots \partial x_k^{\beta_k}},\ x\in\Gamma \quad (8\text{-}139)$$

where

$$\alpha = (\alpha_1, ..., \alpha_k)^T,\ \beta = (\beta_1, ..., \beta_k)^T$$

$$|\alpha| = \alpha_1 + ... + \alpha_k,\ |\beta| = \beta_1 + ... + \beta_k.$$

The object of the problem is to determine $u(x)$, $x\in\Omega$, $v(x)$, $x\in\Gamma$, and b such that

$$J = \int_\Gamma g_0(x,z,v,b)\,d\Gamma + \iint_\Omega f_0(x,z,u,b)\,d\Omega \quad (8\text{-}140)$$

is a minimum subject to the constraints of Eqs. 8-136 and 8-137,

$$\left.\begin{aligned} &\psi_\alpha = \int_\Gamma g_\alpha(x,z,v,b)\,d\Gamma + \iint_\Omega L_\alpha(x,z,u,b)\,d\Omega = 0, \\ &\qquad \alpha = 1, ..., r' \\ &\psi_\alpha = \int_\Gamma g_\alpha(x,z,v,b)\,d\Gamma + \iint_\Omega L_\alpha(x,z,u,b)\,d\Omega < 0, \\ &\qquad \alpha = r'+1, ..., r \end{aligned}\right\} \quad (8\text{-}141)$$

$$\left.\begin{aligned} &\phi_i(x,u) = 0,\ x\in\Omega,\ i = 1, ..., \xi' \\ &\phi_i(x,u) < 0,\ x\in\Omega,\ i = \xi'+1, ..., \xi \end{aligned}\right\} \quad (8\text{-}142)$$

and

$$\left.\begin{aligned} &\omega_j(x,v) = 0,\ x\in\Gamma,\ j = 1, ..., \zeta' \\ &\omega_j(x,v) < 0,\ x\in\Gamma,\ j = \zeta'+1, ..., \zeta. \end{aligned}\right\} \quad (8\text{-}143)$$

The method of solving this problem will be similar to the methods of pars. 8-2 and 8-3. An estimate $u^{(0)}(x)$, $v^{(0)}(x)$, and $b^{(0)}$ will be made and changes sought which reduce J, subject to the constraints of the problem. Before desirable changes in $u^{(0)}$, $v^{(0)}$, and $b^{(0)}$ may be determined, of course, their effect on function in the problem must be examined.

8-4.2 EFFECT OF SMALL CHANGES IN DESIGN VARIABLES AND PARAMETERS

It will be assumed in the following that the boundary-value problem, Eqs. 8-136 and 8-137, is well-behaved in the sense that small changes δu in $u^{(0)}$, δv in $v^{(0)}$, and δb in $b^{(0)}$ yield a new solution $z^{(0)} + \delta z$ (where $z^{(0)}$ is the solution corresponding to the estimated design functions and parameters), where δz is small.

To first order, δz must satisfy the linearized boundary-value problem.

$$L[u^{(0)},b^{(0)}](\delta z) + \Delta_u L[u^{(0)},b^{(0)}][z^{(0)}]\,\delta u + \Delta_b L[u^{(0)},b^{(0)}][z^{(0)}]\,\delta b = \frac{\partial Q}{\partial u}[x,u^{(0)},b^{(0)}]\,\delta u + \frac{\partial Q}{\partial b}[x,u^{(0)},b^{(0)}]\,\delta b \tag{8-144}$$

for $x \in \Omega$ and

$$B[v^{(0)},b^{(0)}](\delta z) + \Delta_v B[v^{(0)},b^{(0)}][z^{(0)}]\,\delta v + \Delta_b B[v^{(0)},b^{(0)}][z^{(0)}]\,\delta b = \frac{\partial q}{\partial v}[x,v^{(0)},b^{(0)}]\,\delta v + \frac{\partial q}{\partial b}[x,v^{(0)},b^{(0)}]\,\delta b \tag{8-145}$$

for $x \in \Gamma$, where

$$\left.\begin{aligned}
\Delta_u L(x,u,b)[z] &= \frac{\partial}{\partial u} L(x,u,b)[z] \\
\Delta_b L(x,u,b)[z] &= \frac{\partial}{\partial b} L(x,u,b)[z] \\
\Delta_v B(x,v,b)[z] &= \frac{\partial}{\partial v} B(x,v,b)[z] \\
\Delta_b B(x,v,b)[z] &= \frac{\partial}{\partial b} B(x,v,b)[z]
\end{aligned}\right\} \tag{8-146}$$

For convenience in the following development, the arguments of L and B will always be taken as $u^{(0)}$, $v^{(0)}$, and $b^{(0)}$.

The functionals J and ψ_α are of the same general form, so, for their analysis define

$$P = \int_\Gamma g(x,z,v,b)\,d\Gamma + \iint_\Omega F(x,z,u,b)\,d\Omega. \tag{8-147}$$

Once the dependence of P on changes in u, v and b is determined, the result may be applied directly to J and ψ_α.

To first order terms,

$$\delta P = \int_\Gamma \left(\frac{\partial g}{\partial z}\,\delta z + \frac{\partial g}{\partial v}\,\delta v + \frac{\partial g}{\partial b}\,\delta b \right) d\Gamma + \iint_\Omega \left(\frac{\partial F}{\partial z}\,\delta z + \frac{\partial F}{\partial u}\,\delta u + \frac{\partial F}{\partial b}\,\delta b \right) d\Omega. \tag{8-148}$$

In order to make use of Eq. 8-148 in the determination of δu, δv, and δb, it is desirable to eliminate explicit dependence on δz. This is done through use of the adjoint operator L^* defined by

$$\iint_{\Omega}\{\lambda^T L(u,b)[\delta z] - \delta z^T L^*(u,b)[\lambda]\}\, d\Omega$$

$$= \int_{\Gamma} A[\lambda]^T C[\delta z]\, d\Gamma \qquad (8\text{-}149)$$

w... $A[\lambda]$ and $C[\delta z]$ are differential operato... The form of the operators A, C, and L^* is determined by integrating $\lambda^T L(u,b)[\delta z]$ by parts.

Putting

$$L^*(u,b)[\lambda] = \frac{\partial F^T}{\partial z} \qquad (8\text{-}150)$$

Eq. 8-148 becomes

$$\delta P = \int_{\Gamma}\left\{\frac{\partial g}{\partial z}\delta z - A[\lambda^T]C[\delta z] + \frac{\partial g}{\partial v}\delta v + \frac{\partial g}{\partial b}\delta b\right\} d\Gamma + \iint_{\Omega}\left\{\lambda^T L(u,b)[\delta z] + \frac{\partial F}{\partial u}\delta u + \frac{\partial F}{\partial b}\delta b\right\} d\Omega\,.$$

Substituting from Eq. 8-144 for $L(u,b)[\delta z]$, this is

$$\delta P = \iint_{\Omega}\left(\left\{\frac{\partial F}{\partial u} - \lambda^T \Delta_u L(u,b)[z] + \lambda^T\frac{\partial Q}{\partial u}\right\}\delta u + \left\{\frac{\partial F}{\partial b} - \lambda^T\Delta_b L(u,b)[z] + \lambda^T\frac{\partial Q}{\partial b}\right\}\delta b\right) d\Omega + \int_{\Gamma}\left\{\frac{\partial g}{\partial z}\delta z - A[\lambda]^T C[\delta z] + \frac{\partial g}{\partial v}\delta v + \frac{\partial g}{\partial b}\delta b\right\} d\Gamma \qquad (8\text{-}151)$$

The objective now is to eliminate explicit dependence of δP on δz. This may be done by requiring that

$$\frac{\partial g}{\partial z}\delta z - A[\lambda]^T C[\delta z] \qquad (8\text{-}152)$$

be explicitly independent of δz for all δz satisfying Eq. 8-145. This may be interpreted as requiring that on Γ certain components of δz be determined from Eq. 8-145 in terms of δv, δb, and the remaining δz. The coefficients of all components of δz remaining in Eq. 8-151 must then be set equal to zero. These equations will then yield boundary conditions for $\lambda(x)$.

Assuming all this calculation has been completed and $\lambda(x)$ determined, Eq. 8-151 may be written as

$$\delta P = \iint_{\Omega}\Lambda^T(x)\delta u\, d\Omega + \int_{\Gamma}\Pi^T(x)\delta v\, d\Gamma + \ell^T\delta b$$

where

$$\Lambda(x) = \frac{\partial F^T}{\partial u} - \Delta_u L(u,b)[z]\lambda + \frac{\partial Q^T}{\partial y}\lambda$$

$\Pi(x)$ = coefficient of δv in Eq. 8-151 after substitution

$$\ell = \iint_{\Omega}\left\{\frac{\partial F^T}{\partial b} - \Delta_b L(u,b)[z]\lambda + \frac{\partial Q^T}{\partial b}\lambda\right\} d\Omega$$

$+ \int_{\Gamma}$ [coefficient of δb in Eq. 8-151 after substitution] $d\Gamma$

(8-153)

By putting

$$g = g_0, \; F = f_0 \tag{8-154}$$

and

$$g = g_\alpha, \; F = L_\alpha \tag{8-155}$$

one obtains

$$\delta J = \iint_\Omega \Lambda^{J^T}(x)\delta u \, d\Omega + \int_\Gamma \Pi^{J^T}(x)\delta v \, d\Omega + \varrho^{J^T}\delta b \tag{8-156}$$

and

$$\delta\psi_\alpha = \iint_\Omega \Lambda^{\psi_\alpha{}^T}(x)\delta u \, d\Omega + \int_\Gamma \Pi^{\psi_\alpha{}^T}(x)\delta v \, d\Gamma + \varrho^{\psi_\alpha{}^T}\delta b \tag{8-157}$$

respectively.

The expressions, Eqs. 8-156 and 8-157, give the desired explicit dependence of δJ and $\delta\psi_\alpha$ on δu, δv, and δb. The problem is now reduced to determination of δu, δv, and δb which give the greatest reduction in J subject to the constraints of the problem

8-4.3 A STEEPEST DESCENT COMPUTATIONAL ALGORITHM

The procedure will now be to choose δu, δv, and δb so as to minimize δJ subject to the constraints Eqs. 8-141 through 8-143, just as in pars 8-1 through 8-3 In order to insure Eq 8-141, define

$$\bar{\psi} = \begin{bmatrix} \psi_\alpha \\ \alpha \in A \end{bmatrix} \tag{8-158}$$

where

$$A = (\alpha \mid \psi_\alpha > 0) . \tag{8-159}$$

It will be required that

$$\delta\bar{\psi} = -C_1\bar{\psi} \tag{8-160}$$

where C_1 is a constant between zero and one. The idea here is to drive ψ_α toward zero if a constraint is violated or will be violated by a change in the design variables or parameters.

For convenience in later development, define

$$\left.\begin{aligned} \Lambda^\psi(x) &= \left(\Lambda^{\psi_\alpha};\ \alpha\in A\right) \\ \Pi^\psi(x) &= \left(\Pi^{\psi_\alpha};\ \alpha\in A\right) \\ \varrho^\psi &= \left(\varrho^{\psi_\alpha},\ \alpha\in A\right). \end{aligned}\right\} \tag{8-161}$$

In this notation,

$$\delta\bar{\psi} = \iint_\Omega \Lambda^\psi(x)^T \delta u \, d\Omega + \int_\Gamma \Pi^{\psi^T}(x)\delta v \, d\Gamma + \varrho^{\psi^T}\delta b . \tag{8-162}$$

Likewise, define

$$\bar{\phi}(x) = \begin{bmatrix} \phi_i \\ i \in D(x) \end{bmatrix} \tag{8-163}$$

and

$$\bar{\omega}(x) = \begin{bmatrix} \omega_j \\ j \in E(x) \end{bmatrix} \qquad (8\text{-}164)$$

where

$$D(x) = \{i \mid \phi_i(x) > 0\} \qquad (8\text{-}165)$$

and

$$E(x) = \{j \mid \omega_j(x) > 0\} \qquad (8\text{-}166)$$

It will be required that

$$\delta\bar{\phi}(x) < -C_2\bar{\phi}(x), \quad x \in \Omega \qquad (8\text{-}167)$$

and

$$\delta\bar{\omega}(x) < -C_3\bar{\omega}(x), \quad x \in \Gamma \qquad (8\text{-}168)$$

where

$0 < C_2 < 1$ and $0 < C_3 < 1$

$$\delta\bar{\phi}(x) = \frac{\partial\bar{\phi}}{\partial u}(x)\delta u(x) \qquad (8\text{-}169)$$

and

$$\delta\bar{\omega}(x) = \frac{\partial\bar{\omega}}{\partial v}(x)\delta v(x). \qquad (8\text{-}170)$$

Before determining δu, δv, and δb, a device should be introduced to insure that these quantities are small as is required in order that the preceding first order analysis is a good approximation to reality. The engineer should choose positive definite weighting matrices $W_u(x)$, $W_v(x)$, and W_b so as to associate a relative importance to all the variables. It is then required that

$$dP^2 = \iint_\Omega \delta u^T W_u \delta u \, d\Omega + \int_\Gamma \delta v^T W_v \delta v \, d\Gamma + \delta b^T W_b \delta b \qquad (8\text{-}171)$$

where dP is "small". The choice of dP will be discussed later.

The design variables and parameter, δu, δv and δb, are now to be chosen to minimize δJ of Eq. 8-156 subject to Eqs. 8-160, 8-167, 8-168, and 8-171.

A multiplier rule of Liusternik and Sobolev, Ref. 14, page 209, will now be applied to the present problem. It guarantees the existence of multipliers, $\mu(x), x \in \Omega$, $\mu_i(x) > 0$, $i > \xi'$, $\nu(x), x \in \Gamma$, $\nu_j(x) > 0$, $j > \zeta'$, γ, $\gamma_\alpha > 0$, $\alpha > r'$, $\lambda_0 > 0$, and γ_0 such that

$$\delta(\delta\bar{J}) = 0 \qquad (8\text{-}172)$$

for all δu, δv and δb, where

$$\delta\bar{J} = \iint_\Omega \Big[-\lambda_0 \Lambda^{J^T}(x) - \gamma^T \Lambda^{\psi^T}(x) - \gamma_0 \delta u^T W_u - \mu^T(x)\frac{\partial\bar{\phi}}{\partial u} \Big] \delta u \, d\Omega + \int_\Gamma \Big[-\lambda^0 \Pi^{J^T}(x) - \gamma^T \Pi^{\psi^T}(x) - \gamma_0 \delta v^T W_v - \nu^T(x)\frac{\partial\bar{\omega}}{\partial v} \Big] \delta v \, d\Gamma + \Big[-\lambda_0 \ell^{J^T} \quad \gamma^T \ell^{\psi^T} - \gamma_0 \, \delta b \, W_b \Big] \delta b \qquad (8\text{-}173)$$

Using $\delta\bar{J}$ of Eq. 8-173 in Eq. 8-172

$$
\begin{aligned}
\delta(\delta\bar{J}) = 0 = \iint_{\Omega} \Big[& -\lambda_0 \Lambda^{J^T}(x) \\
& - \gamma^T \Lambda^{\psi^T}(x) \\
& - 2\gamma_0 \delta u^T W_u \\
& - \mu^T(x) \frac{\partial \bar{\phi}}{\partial u} \Big] \delta^2 u \, d\Omega \\
+ \int_{\Gamma} \Big[& -\lambda_0 \Pi^{J^T}(x) \\
& - \gamma^T \Pi^{\psi^T}(x) - 2\gamma_0 \delta \nu^T W_\nu \\
& - \nu^T(x) \frac{\partial \bar{\omega}}{\partial \nu} \Big] \delta^2 \nu \, d\Gamma \\
+ \Big[& -\lambda_0 \varrho^{J^T} - \gamma^T \varrho^{\psi} \\
& - 2\gamma_0 \delta b W_b \Big] \delta^2 b \, ,
\end{aligned}
\tag{8-174}
$$

for all $\delta^2 u(x)$, $x \in \Omega$, $\delta^2 \nu(x)$, $x \in \Gamma$, and $\delta^2 b$. This implies

$$
\begin{aligned}
& -\lambda_0 {}^J(x) - \Lambda^{\psi}(x) - 2\gamma_0 W_u \delta u(x) \\
& - \frac{\partial \bar{\phi}^T}{\partial u} \mu(x) = 0
\end{aligned}
\tag{8-175}
$$

for $x \in \Omega$

$$
\begin{aligned}
& -\lambda_0 \Pi^J(x) - \Pi^{\psi}(x)\gamma - 2\gamma_0 W_\nu \delta \nu(x) \\
& - \frac{\partial \bar{\omega}^T}{\partial \nu} \nu(x) = 0
\end{aligned}
\tag{8-176}
$$

for $x \in \Gamma$, and

$$
-\lambda_0 \varrho^J - \varrho^{\psi} \gamma - 2\gamma_0 W_b b = 0 \, . \tag{8-177}
$$

At this point it is assumed that the problem is normal so that $\lambda_0 = 1$ may be chosen. Eqs. 8-175 through 8-177 yield

$$
\begin{aligned}
\delta u(x) = \frac{1}{2\gamma_0} & W_u^{-1}(x) \\
& \times \left[-\Lambda^J(x) - \Lambda^{\psi}(x)\gamma - \frac{\partial \bar{\phi}^T}{\partial u} \mu(x) \right] , \\
& x \in \Omega
\end{aligned}
\tag{8-178}
$$

$$
\begin{aligned}
\delta \nu(x) = \frac{1}{2\gamma_0} & W_\nu^{-1}(x) \\
& \times \Big[-\Pi^J(x) - \Pi^{\psi}\gamma \\
& \quad - \frac{\partial \bar{\omega}^T}{\partial \nu} \nu(x) \Big] , x \in \Gamma
\end{aligned}
\tag{8-179}
$$

and

$$
\delta b = \frac{1}{2\gamma_0} W_b^{-1} \left(-\varrho^J - \varrho^{\psi} \gamma \right) \tag{8-180}
$$

Assume for the present that Eqs. 8-167 and 8-168 are equalities. Substituting Eqs. 8-178 and 8-179 into Eqs. 8-167 and 8-168 yields

$$
\begin{aligned}
& \frac{1}{2\gamma_0} \frac{\partial \bar{\phi}}{\partial u} W_u^{-1} \left(-\Lambda^J - \Lambda^{\psi}\gamma - \frac{\partial \bar{\phi}^T}{\partial u} \mu \right) = \\
& \qquad C_2 \bar{\phi}, \; x \in \Omega
\end{aligned}
$$

and

$$
\begin{aligned}
& \frac{1}{2\gamma_0} \frac{\partial \bar{\omega}}{\partial \nu} W_\nu^{-1} \Big(-\Pi^J - \Pi^{\psi}\gamma \\
& \qquad - \frac{\partial \bar{\omega}^T}{\partial \nu} \nu \Big) = C_3 \bar{\omega} \; x \in \Gamma
\end{aligned}
$$

Since W_u^{-1} and W_v^{-1} are positive definite, the matrices $(\partial\bar{\phi}/\partial u)\ W_u^{-1}\ (\partial\bar{\phi}^T/\partial u)$ and $(\partial\bar{\omega}/\partial v)\ W_v^{-1}\ (\partial\bar{\omega}^T/\partial v)$ are positive semi-definite. It will be assumed that they are positive definite and hence nonsingular. In case $\bar{\phi}$ or $\bar{\omega}$ is empty, then the terms multiplying μ and ν do not exist. In this case simply define $\mu = \nu = 0$ and $(\partial\bar{\phi}/\partial u)\ W_u^{-1}\ (\partial\bar{\phi}^T/\partial u) = (\partial\bar{\omega}/\partial v)\ W_v^{-1}\ (\partial\bar{\omega}^T/\partial v) = I$. In any case,

$$\mu(x) = \Lambda^{\phi^{-1}}\left[2\gamma_0 C_2\bar{\phi} + \frac{\partial\bar{\phi}}{\partial u} W_u^{-1} \times (-\Lambda^J - \Lambda^\psi\gamma)\right],\ x\in\Omega \tag{8-181}$$

where

$$\Lambda^\phi = \frac{\partial\bar{\phi}}{\partial u} W_u^{-1} \frac{\partial\bar{\phi}^T}{\partial u}$$

and

$$\nu(x) = \Lambda^{\omega^{-1}}\left[2\gamma_0 C_3\bar{\omega} + \frac{\partial\bar{\omega}}{\partial v} W_v^{-1} (-\Pi^J - \Pi^\psi\gamma)\right],\ x\in\Gamma \tag{8-182}$$

where

$$\Lambda^\omega = \frac{\partial\bar{\omega}}{\partial v} W_v^{-1} \frac{\partial\bar{\omega}^T}{\partial v}.$$

Substituting from Eqs. 8-181 and 8-182 into Eqs. 8-178 and 8-179 yields

$$\delta u(x) = \frac{1}{2\gamma_0} W_u^{-1}\left[\left(I - \frac{\partial\bar{\phi}^T}{\partial u}\Lambda^{\phi^{-1}}\frac{\partial\bar{\phi}}{\partial u}W_u^{-1}\right) \times (-\Lambda^J - \Lambda^\psi\gamma) - \frac{\partial\bar{\phi}^T}{\partial u}\Lambda^{\phi^{-1}}2\gamma_0 C_2\bar{\phi}\right],\ x\in\Omega \tag{8-183}$$

$$\delta v(x) = \frac{1}{2\gamma_0} W_v^{-1}\left[\left(I - \frac{\partial\bar{\omega}^T}{\partial v}\Lambda^{\omega^{-1}}\frac{\partial\bar{\omega}}{\partial v}W_v^{-1}\right) \times (-\Pi^J - \Pi^\psi\gamma) - \frac{\partial\bar{\omega}^T}{\partial v}\Lambda^{\omega^{-1}}2\gamma_0 C_3\bar{\omega}\right],\ x\in\Gamma. \tag{8-184}$$

In order to determine γ, these expressions are substituted into Eq. 8-160. Using Eq. 8-182, the resulting equation is

$$-\frac{1}{2\gamma_0}M_{\psi J} - \frac{1}{2\gamma_0}M_{\psi\psi}\gamma - C_2 M_{\psi\bar{\phi}} - C_3 M_{\psi\bar{\omega}} = -C_1\bar{\psi} \tag{8-185}$$

where

$$M_{\psi J} = \iint_\Omega \Lambda^{\psi^T} W_u^{-1}\left(I - \frac{\partial\bar{\phi}^T}{\partial u}\Lambda^{\phi^{-1}} \times \frac{\partial\bar{\phi}}{\partial u}W_u^{-1}\right)\Lambda^J d\Omega + \int_\Gamma \Pi^{\psi^T} W_v^{-1}\left(I - \frac{\partial\bar{\omega}^T}{\partial v}\Lambda^{\omega^{-1}} \times \frac{\partial\bar{\omega}}{\partial v}W_v^{-1}\right)\Pi^J d\Gamma + \ell^{\psi^T} W_b^{-1}\ell^J \tag{8-186}$$

$$M_{\psi\psi} = \iint_{\Omega} \Lambda^{\psi T} W_u^{-1} \left(I - \frac{\partial \bar{\phi}^T}{\partial u} \Lambda^{\phi^{-1}} \times \frac{\partial \bar{\phi}}{\partial u} W_u^{-1} \right) \Lambda^{\psi} d\Omega$$

$$+ \int_{\Gamma} \Pi^{\psi T} W_v^{-1} \left(I - \frac{\partial \bar{\omega}^T}{\partial v} \Lambda^{\psi^{-1}} \times \frac{\partial \bar{\omega}}{\partial v} W_v^{-1} \right) \Pi^{\psi}$$

$$+ \ell^{\psi T} W_b^{-1} \ell^{\psi} \qquad (8\text{-}187)$$

$$M_{\psi\bar{\phi}} = \iint_{\Omega} \Lambda^{\psi T} W_u^{-1} \frac{\partial \bar{\phi}^T}{\partial u} \Lambda^{\phi^{-1}} \bar{\phi} \, d\Omega \qquad (8\text{-}188)$$

and

$$M_{\psi\bar{\omega}} = \int_{\Gamma} \Pi^{\psi T} W_v^{-1} \frac{\partial \bar{\omega}^T}{\partial v} \Lambda^{\omega^{-1}} \bar{\omega} \, d\Gamma. \qquad (8\text{-}189)$$

It was shown in par. 8-2 that the matrices in the integrands for $M_{\psi\psi}$ are positive semi-definite. Therefore, $M_{\psi\psi}$ is at least positive semi-definite. It will be assumed in what follows that $M_{\psi\psi}$ is positive definite and, therefore, nonsingular. Solving Eq. 8-185 for γ then yields

$$\gamma = M_{\psi\psi}^{-1} \left[2\gamma_0 (C_1 \bar{\psi} - C_2 M_{\psi\bar{\phi}} - C_3 M_{\psi\bar{\omega}}) - M_{\psi J} \right]. \qquad (8\text{-}190)$$

Substituting γ from Eq. 8-190 into Eq 8-180, Eqs. 8-183, and 8-184 yield

$$\delta u(x) = -\frac{1}{2\gamma_0} \delta u^1(x) + \delta u^2(x) \qquad (8\text{-}191)$$

$$\delta v(x) = -\frac{1}{2\gamma_0} \delta v^1(x) + \delta v^2(x) \qquad (8\text{-}192)$$

and

$$\delta b = -\frac{1}{2\gamma_0} \delta b^1 + \delta b^2 \qquad (8\text{-}193)$$

where

$$\delta u^1(x) = W_u^{-1} \left(I - \frac{\partial \bar{\phi}^T}{\partial u} \Lambda^{\phi^{-1}} \frac{\partial \bar{\phi}}{\partial u} W_u^{-1} \right) \times \left(\Lambda^J - \Lambda^{\psi} M_{\psi\psi}^{-1} M_{\psi J} \right), x \in \Omega \qquad (8\text{-}194)$$

$$\delta u^2(x) = W_u^{-1} \left(I - \frac{\partial \bar{\phi}^T}{\partial u} \Lambda^{\phi^{-1}} \frac{\partial \bar{\phi}}{\partial u} W_u^{-1} \right) \times \left[\Lambda^{\psi} M_{\psi\psi}^{-1} (-C_1 \bar{\psi} + C_2 M_{\psi\bar{\phi}} + C_3 M_{\psi\bar{\omega}}) \right] \qquad (8\text{-}195)$$

$$- C_2 W_u^{-1} \frac{\partial \bar{\phi}^T}{\partial u} \Lambda^{\phi^{-1}} \bar{\phi}, \; x \in \Omega$$

$$\delta v^1(x) = W_v^{-1} \left(I - \frac{\partial \bar{\omega}^T}{\partial v} \Lambda^{\omega^{-1}} \frac{\partial \bar{\omega}}{\partial v} W_v^{-1} \right) \times \left(\Pi^J - \Pi^{\psi} M_{\psi\psi}^{-1} M_{\psi J} \right), x \in \Gamma \qquad (8\text{-}196)$$

$$\delta v^2(x) = W_v^{-1} \left(I - \frac{\partial \bar{\omega}^T}{\partial v} \Lambda^{\omega^{-1}} \frac{\partial \bar{\omega}}{\partial v} W_v^{-1} \right) \times \left[\Pi^{\psi} M_{\psi\psi}^{-1} (C_1 \bar{\psi} + C_2 M_{\psi\bar{\phi}} + C_3 M_{\psi\bar{\omega}}) \right] \qquad (8\text{-}197)$$

$$C_3 W_v^{-1} \frac{\partial \bar{\omega}^T}{\partial v} \Lambda^{\omega^{-1}} \bar{\omega}, \; x \in \Gamma$$

$$\delta b^1 = W_b^{-1}\left(\ell^J - \ell^\psi M_{\psi\psi}^{-1} M_{\psi J}\right) \tag{8-198}$$

and

$$\delta b^2 = W_b^{-1} \ell^\psi M_{\psi\psi}^{-1}\left(-C_1\psi + C_2 M_{\psi\bar\phi} + C_3 M_{\psi\bar\omega}\right). \tag{8-199}$$

It should be noted that if there were no constraints then δu, δv, and δb would reduce to $-\frac{1}{2\gamma_0}\Lambda^J$, $-\frac{1}{2\gamma_0}\Pi^J$, and $-\frac{1}{2\gamma_0}\ell^J$, respectively. In order that the change in design variables and parameters should be in the negative gradient direction, it is clear that $\gamma_0 > 0$ is required. The magnitude of γ_0 could be determined by substituting Eqs. 8-191 through 8-199 into Eq. 8-171. However, dP must be chosen so it may be just as well to simply choose γ_0 in Eqs. 8-191, 8-192 and 8-193.

Just as in the problems of pars. 8-2 and 8-3, the variations $\delta u^1(x)$, $\delta u^2(x)$, $\delta v^1(x)$, $\delta v^2(x)$, δb^1, and δb^2 satisfy Theorem, 8-2.

Theorem 8-2: The above variations satisfy the identities

1. $\delta b^{1T} W_b \delta b^2 + \int_\Gamma \delta v^1 W_v \delta v^2 d\Gamma + \iint_\Omega \delta u^1 W_u \delta u^2 d\Omega = 0$

2. $\ell^{\psi T}\delta b^2 + \int_\Gamma \Pi^{\psi T}\delta v^2 d\Gamma + \iint_\Omega \Lambda^{\psi T}\delta u^2 d\Omega = -C_1\bar\psi$

3. $\ell^{\psi T}\delta b^1 + \int_\Gamma \Pi^{\psi T}\delta v^1 d\Gamma + \iint_\Omega \Lambda^{\psi T}\delta u^1 d\Omega = 0$

4. $\frac{\partial\bar\phi}{\partial u}\delta u^2 = -C_2\bar\phi$, in Ω

5. $\frac{\partial\bar\phi}{\partial u}\delta u^1 = 0$, in Ω

6. $\frac{\partial\bar\omega}{\partial v}\delta v^2 = -C_3\bar\omega$, on Γ

7. $\frac{\partial\bar\omega}{\partial v}\delta v^1 = 0$, on Γ

8. $-\ell^{J^T}\delta b^1 - \int_\Gamma \Pi^{J^T}\delta v^1 d\Gamma = \iint_\Omega \Lambda^{J^T}\delta u^1 d\Omega < 0$

A computational algorithm may now be stated based on this development and the arguments presented in par. 8-2.

Algorithm

Step 1. Make an engineering estimate $u^{(0)}(x)$, $v^{(0)}(x)$, $b^{(0)}$ of the optimum design functions and parameter.

Step 2. Solve Eqs. 8-136 and 8-137 for $z^{(0)}(x)$ corresponding to $u^{(0)}(x)$, $v^{(0)}(x)$, and $b^{(0)}$.

Step 3. Check constraints and form $\bar\psi$, $\bar\phi$, and $\bar\omega$ of Eqs. 8-158, 8-163, and 8-164.

Step 4. Solve the differential equation, Eq. 8-150, subject to the boundary conditions generated by Eq. 8-152 with g and F given by Eqs. 8-154 and 8-155, to obtain λ^J and $\lambda^{\psi\alpha}$, respectively.

Step 5. Compute $\Lambda^J(x)$, $\Pi^J(x)$, ℓ^J, $\Lambda^\psi(x)$,

$\Pi^{\psi}(x)$, and ℓ^{ψ} in Eqs. 8-153 and 8-161.

Step 6. Choose the correction factors C_1, C_2, and C_3 in Eqs. 8-160, 8-167, and 8-168.

Step 7. Compute $M_{\psi J}$, $M_{\psi\psi}$, $M_{\psi\tilde{\phi}}$, and $M_{\psi\tilde{\omega}}$ in Eqs. 8-186 through 8-189.

Step 8. Choose $\gamma_0 > 0$ and compute γ, $\mu(x)$, and $\nu(x)$ in Eqs. 8-190, 8-181, and 8-182. If any components of γ with $\alpha > r'$, $\mu(x)$ with $i > \xi'$, or $\nu(x)$ with $j > \zeta'$ are negative, re-define $\tilde{\psi}$, $\tilde{\phi}(x)$, and $\tilde{\omega}(x)$ by deleting corresponding terms and return to Step 5.

Step 9. Compute $\delta u^1(x)$, $\delta u^2(x)$, $\delta v^1(x)$, $\delta v^2(x)$, δb^1, and δb^2 in Eqs. 8-194 through 8-199.

Step 10. Compute

$$u^{(1)}(x) = u^{(0)}(x) - \frac{1}{2\gamma_0}\delta u^1(x) + \delta u^2(x)$$

$$v^{(1)}(x) = v^{(0)}(x) - \frac{1}{2\gamma_0}\delta v^1(x) + \delta v^2(x)$$

$$b^{(1)} = b^{(0)} - \frac{1}{2\gamma_0}\delta b^1 + \delta b^2 .$$

Step 11. If the constraints are satisfied and $\delta u^1(x)$, $\delta v^1(x)$, and δb^1 are sufficiently small, terminate. Otherwise, return to Step 2 with $u^{(0)}(x)$, $v^{(0)}(x)$, and $b^{(0)}$ replaced by $u^{(1)}(x)$, $v^{(1)}(x)$, and $b^{(1)}$, respectively.

As in par. 8-2, if after several iterations the constraints are all satisfied, $\delta u^2(x)$, $\delta v^2(x)$, and δb^2 will all be zero. In this case,

$$\delta J = -\frac{1}{2\gamma_0}\left(M_{JJ} - M^T_{\psi J} M^{-1}_{\psi\psi} M_{\psi J}\right) \quad (8\text{-}200)$$

where

$$M_{JJ} = \ell^{J^T} W_b \ell^J + \int_\Gamma \Pi^{J^T} W_v^{-1} \times \left(I - \frac{\partial\tilde{\omega}}{\partial v}\Lambda^{\omega^{-1}}\frac{\partial\tilde{\omega}}{\partial v}W_v\right)\Pi^J d\Gamma + \iint_\Omega \Lambda^{J^T} W_u^{-1} \times \left(I - \frac{\partial\tilde{\phi}}{\partial u}\Lambda^{\phi^{-1}}\frac{\partial\tilde{\phi}}{\partial u}W_u\right)\Lambda^J d\Omega .$$

Just as in par. 8-2, one can now specify a reasonable, desired reduction δJ in J. The formulation, Eq. 8-200, provides a means of finding γ_0 that, based on the preceding linear approximations, will yield the desired reduction in the cost function.

8-5 OPTIMAL DESIGN OF AN ARTILLERY RECOIL MECHANISM*

An artillery weapon mounted on tires or tracks has some undesirable features. Unlike the hard mount (weapon rests on a base plate), the flexible mount will have a pitch motion. During the recoil stroke, when the weapon is fired at 75-deg elevation, the tires load up or compress; and when counterrecoil begins, the tires act like a spring and unload sending the tires off the ground. It is quite obvious that, when the weapon comes to rest, the likelihood of it being zeroed in for the

*The results of this paragraph represent the work of Mr. T D Streeter, Ref 15.

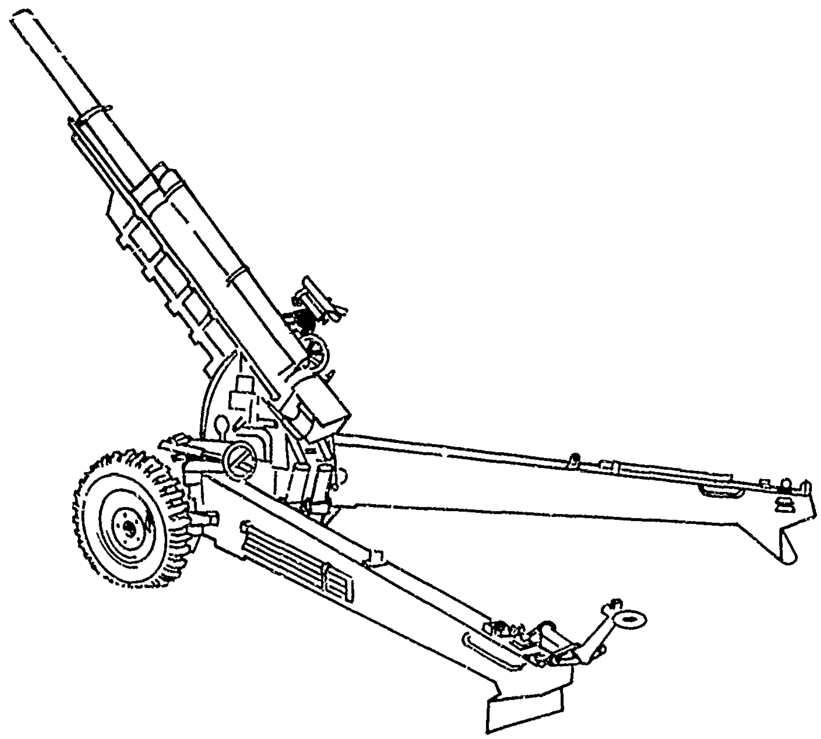


Figure 8-1. Howitzer, Towed, 105 mm, XM164

next round has been reduced considerably, especially for high rate of fire weapons. This phenomenon is known as a secondary recoil effect, because an additional acceleration term enters into the recoil equations. Because of this secondary recoil effect, the recoil mechanism design becomes much more difficult. For short recoil, the orifice areas in the recoil mechanism are designed at maximum elevation (75 deg). Therefore, when elevation is mentioned throughout the remainder of this report, it refers to maximum elevation. The weapon positioned for high-angle fire is shown in Fig. 8-1.

The purpose of this paragraph will be to develop a systematic recoil mechanism design procedure characterized by mathematical modeling for the high-speed digital computer. To do this, the steepest descent numerical technique is used to minimize the hop or pitch motion of the weapon and, at the same time, to determine the necessary control rod * design that will minimize hop

The recoil equation for a rigid mount is of the form

$$\ddot{x} + f(x)\dot{x}^2 + g(x) = h(t) \qquad (8\text{-}201)$$

where x is the displacement of the recoiling parts, $g(x)$ is a restoring force, and $h(t)$ is the

The control rod in a hydraulic recoil mechanism is a rod of variable cross section which moves through a larger orifice during recoil and varies the area of the orifice to control recoil force level breech force due to firing. In the second term of this equation, the expression for the effect of the control rod orifice areas can also be obtained from a predetermined recoil force, $R(t)$. For the flexible mount, Eq. 8-201 is coupled with the equation describing the pitch motion of the weapon and thus yielding two second-order nonlinear ordinary differential equations with prescribed initial conditions. $R(t)$ will be taken as the control variable which is to be determined to minimize hop (the pitch motion of the weapon) subject to other design constraints. The orifice area is then determined to provide this recoil force.

This study was performed on a developmental weapon, namely, the XM164. The XM164 is a lightweight, split-trailed towed 105 mm howitzer with the XM44 hydropneumatic recoil mechanism. Unlike a rigid mount, the XM164 is flexible and is fired while resting on rubber tires.

For a rigid mount weapon, the resisting force $R(t)$ on the recoiling parts is designed with a trapezoidal shape as shown in Fig. 8-2. With the proper design of the control rod orifice area, the flow of oil in the recoil mechanism is controlled and such a force, as shown in Fig. 8-2, can be obtained. However, when a force (shaped as in Fig. 8-2) is designed for the flexible mount, the question is asked, "Can this force be applied with some other 'best' shape, such that it will reduce the pitch of the weapon?" This is the basic question with which this design problem is concerned.

*The recoil force is the retarding force on the rearward traveling barrel during recoil, due to throttling of oil through the variable area orifice

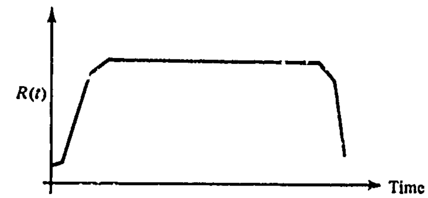


Figure 8-2. Recoil Force for a Rigid Mount

8-5.1 FORMULATION OF THE PROBLEM

During the recoil, counterrecoil cycle, there are four different times which are of concern. These are shown in Fig. 8-3.

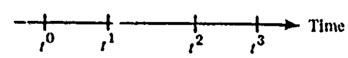


Figure 8-3. Time Intervals

In Fig. 8-3, the special times noted are:

t^0 = firing of round

t^1 = end of the recoil stroke

t^2 = time at which maximum hop occurs

t^3 = end of counterrecoil

At these four times certain conditions must be satisfied from the design requirements. At time t^0 the initial conditions for the state of the system are given. At time t^1 the displacement of the recoiling parts is required to be equal to some specified value and the velocity of the recoiling parts must be equal to zero.

At time t^2 the velocity of the pitch motion must be zero (necessary condition for maximum pitch), and the displacement of the pitch motion is to be a minimum. Note that it will be possible for t^2 to vary between t^1 and t^3. Therefore, the hop or pitch motion will be minimized for the entire counterrecoil stroke. At the final time t^3, which is the end of counterrecoil, the recoiling parts must return to their original position, and the velocity of the recoiling parts will be some specified value V_3. This is to insure that the recoiling parts return to the latch position. It will also be demanded that the total cycle time be equal to c_T seconds.

Formulating this problem into the mathematical notation of par. 8-3 yields

$$\text{Minimize } J = x_4(t^2) \equiv g_0 \tag{8-202}$$

subject to the equality constraints:

$$\left.\begin{aligned}
\psi_1 &= x_2(t^1) - \eta_0 + \eta_{max} \equiv g_1 = 0\\
\psi_2 &= x_2(t^3) - \eta_0 \equiv g_2 = 0\\
\psi_3 &= x_1(t^3) - V_3 \equiv g_3 = 0\\
\Omega^1 &= x_1(t^1) = 0\\
\Omega^2 &= x_3(t^2) = 0\\
\Omega^3 &= t^3 - c_T = 0
\end{aligned}\right\} \tag{8-203}$$

with the full set of initial conditions

$$\left.\begin{aligned}
&x_1(0) = x_3(0) = x_4(0) = 0,\\
&x_2(0) = \eta_0
\end{aligned}\right\} \tag{8-204}$$

where ψ_i, i = 1,2,3 are intermediate and terminal constraint functions to be satisfied; Ω^1, Ω^2, and Ω^3 define the times at which the intermediate and terminal constraint functions occur; x_1 and x_3 are the velocities of the recoiling parts and pitch motion, respectively; x_2 and x_4 are the displacements of the recoiling parts and pitch motion, respectively; $\psi_1 = 0$ is the constraint on the displacement of the recoiling parts such that at the end of the recoil stroke the displacement will be exactly equal to η_{max} inches. $\psi_2 = 0$ is the constraint demanding that the recoiling parts return to the latch position at the end of counterrecoil. $\psi_3 = 0$ is the constraint which requires that the velocity of the recoiling parts come into the latch position at a velocity V_3 inches per second. $\Omega^1 = 0$ defines the time at which the end of the recoil occurs; $\Omega^2 = 0$ defines the times at which the pitch velocity is zero and the one with the largest displacement is selected, thus defining the time at which maximum hop occurs; and $\Omega^3 = 0$ defines the total cycle time to be exactly equal to c_T seconds.

It was previously mentioned that the rod force was taken as the design (control) variable, instead of the orifice areas. Using the rod force as the design variable simplifies the problem, and it also gives the engineer more insight into the design process since he has an intuitive feel for the force levels the weapon system he is designing can tolerate. Thus, immediately the engineer can specify an admissible upper limit for the recoil force, say R_{max}, for his design, and this value may be varied by the engineer for any redesign. The following inequality constraint, therefore, must hold for all time t.

$$\phi = R(t) - R_{max} < 0 \quad 0 < t < t^3 \tag{8-205}$$

The optimization problem has now been formulated. All that must be done now is to put the problem into the steepest descent formulation. Par. 8-5.2 simplifies the equations of motion for the XM164 Howitzer.

8-5.2 EQUATIONS OF MOTION FOR THE XM164 HOWITZER

A schematic diagram showing the moving parts and variables defining the dynamic model of the XM164 Howitzer is shown in Fig. 8-4. In explanation of this physical idealization, the following variables are defined:

W_a = weight of recoiling parts

W_b = weight of elevating parts less recoiling parts

W_d = weight of nonelevating parts

M_a = mass of recoiling parts

η = recoil displacement

ζ = distance from center line of trunnion to center line of recoiling parts (vertical)

y_t = distance from center line of spade to center line of trunnion (horizontal)

γ = angle of elevation of gun tube

z_t = distance from center line of spade to center line of trunnion (vertical)

ϕ = pitch angle of weapon

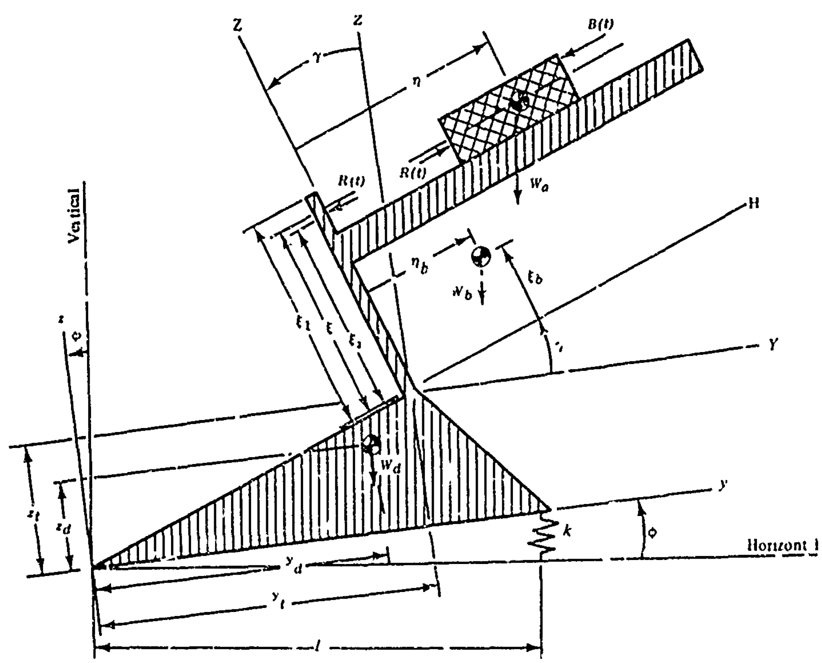


Figure 8-4. Schematic of XM164 105 mm Towed Howitzer – Dynamic Model

$R(t)$ = rod force

$B(t)$ = breech force

g = acceleration due to gravity

μ = coefficient of friction

S_1 = friction force (guide)

S_2 = friction force (guide)

M_b = mass of elevating parts less recoiling parts

η_b = distance from center line of trunnion to mass center of W_b (horizontal)

ζ_b = distance from center line of trunnion to mass center of W_b (vertical)

M_d = mass of nonelevating parts

y_d = distance from center line of spade to mass center of W_d (horizontal)

z_d = distance from center line of spade to mass center of W_d (vertical)

I_a = traverse moment of inertia of W_a about its own CG

I_b = traverse moment of inertia of W_b about its own CG

I_d = traverse moment of inertia of W_d about its own CG

ζ_1 = distance from center line of trunnion to center line of $B(t)$ application

ϕ_{st} = static value of ϕ

c = damping coefficient

k = spring constant of tire

q_1 = distance from center line of trunnion to rear of cradle (horizontal)

q_2 = distance from center line of trunnion to front of cradle (horizontal)

ζ_2 = distance from center line of trunnion to $R(t)$ application (vertical)

α = distance from center line of trunnion to bottom of rail (vertical)

β = distance from center line of trunnion to top of rail (vertical)

L = distance from tires to trail spades

Y,Z = axes fixed in trunnion, parallel to x-, y-axes

y,z = axes fixed in carriage

H,Z = axes fixed in cradle

The differential equations to be solved are, Ref. 16:

$$M_a[\ddot{\eta} - (\zeta - y_t \sin\gamma + z_t \cos\gamma)\ddot{\phi} - (\eta + y_t \cos\gamma + z_t \sin\gamma)\dot{\phi}^2] = R(t) \quad B(t) - M_a g \sin(\gamma + \phi)$$

$$-\mu(|S_1| + |S_2|)\,\mathrm{sgn}(\dot{\eta}) \qquad (8\text{-}206)$$

$$\Big\{M_a(\eta + y_t\cos\gamma + z_t\sin\gamma)^2 + M_b[(\eta_b\cos\gamma + y_t - \zeta_b\sin\gamma)^2 + (\eta_b\sin\gamma + z_t + \zeta_b\cos\gamma)^2] + M_d(y_d^2 + z_d^2) + I_a + I_b + I_d\Big\}\ddot{\phi} + 2M_a\dot{\eta}\dot{\phi}(\eta + y_t\cos\gamma + z_t\sin\gamma)$$

$$- M_a\dot{\phi}^2(\eta + y_t\cos\gamma + z_t\sin\gamma) \times (\zeta - y_t\sin\gamma + z_t\cos\gamma) = B(t)\cdot(\zeta_1 - \zeta)$$

$$+ [R(t) - \mu(|S_1| + |S_2|)\,\mathrm{sgn}(\dot{\eta})] \times (\zeta - y_t\sin\gamma + z_t\cos\gamma)$$

$$- g\Big\{M_a(\eta + y_t\cos\gamma + z_t\sin\gamma) \times \cos(\gamma + \phi) + M_d(y_d\cos\phi - z_d\sin\phi)$$

$$+ M_b[y_t\cos\phi - z_t\sin\phi + \eta_b\cos(\gamma+\phi) - \zeta_b\sin(\gamma+\phi)]\Big\} - k(\phi + \phi_{st}) - c\dot{\phi}$$

$$(8\text{-}207)$$

$$M_a[2\dot{\eta}\dot{\phi} + (\eta + y_t\cos\gamma + z_t\sin\gamma)\ddot{\phi} - (\zeta + z_t\cos\gamma - y_t\sin\gamma)\dot{\phi}^2]$$

$$= S_1 + S_2 - M_a g\cos(\gamma + \phi) \qquad (8\text{-}208)$$

$$I_a\ddot{\phi} = S_1(q_1 - \eta) + S_2(q_2 - \eta) - B(t)\cdot(\zeta - \zeta_1) + R(t)\cdot(\zeta - \zeta_2)$$

$$- \mu[|S_1|(\zeta - \alpha) + |S_2|(\zeta - \beta)]\,\mathrm{sgn}(\dot{\eta})\,. \qquad (8\text{-}209)$$

Eqs. 8-206 and 8-207 are the translational and rotational equations of motion, respectively, for the XM164 Howitzer. Eqs. 8-208 and 8-209 determine the guide friction.

For small ϕ the following approximations are made:

$$\sin\phi = \phi$$

$$\cos\phi = 1 - \frac{\phi^2}{2}$$

The cos $(\gamma + \phi)$ and the sin $(\gamma + \phi)$ then become

$$\cos(\gamma + \phi) = \cos\gamma - \phi^2\frac{\cos\gamma}{2} - \phi\sin\gamma$$

$$\sin(\gamma + \phi) = \sin\gamma - \phi^2\sin\gamma/2 + \phi\cos\gamma\,.$$

From these approximations and the following definitions, Eqs. 8-206 and 8-207 can be simplified. For simplification, define:

$\mathrm{CON1} = M_a$

$\mathrm{CON2} = -M_a(\zeta - y_t\sin\gamma + z_t\cos\gamma)$

$\mathrm{CON3} = -\mu(|S_1| + |S_2|)\,\mathrm{sgn}(\dot{\eta})$

$\mathrm{CON4} = -y_t\cos\gamma + z_t\sin\gamma$

$$\mathrm{CON5} = M_b\Big[(\eta_b\cos\gamma + y_t - \zeta_b\sin\gamma)^2 + (\eta_b\sin\gamma + z_t + \zeta_b\cos\gamma)^2\Big] + M_d(y_d^2 + z_d^2) + I_a + I_b + I_d$$

$\mathrm{CON6} = \zeta - y_t\sin\gamma + z_t\cos\gamma$

$\mathrm{CON7} = \zeta_1 - \zeta$

$\mathrm{CON8} = \sin\gamma$

$\mathrm{CON9} = \cos\gamma$

$\mathrm{CON10} = -M_a\cdot g\cdot\mathrm{CON8}$

$\mathrm{CON11} = M_a\cdot g\cdot\mathrm{CON8}/2$

$$CON12 = M_a \cdot CON4$$

$$CON13 = -2M_a$$

$$CON14 = -2M_a \cdot CON4$$

$$CON15 = -k \cdot \phi_{st}$$

$$CON16 = M_a \cdot CON6$$

$$CON17 = M_a \cdot CON6 \cdot CON4$$

$$CON18 = -g \cdot M_a \cdot CON9$$

$$CON19 = g \cdot M_a \cdot CON9/2$$

$$CON20 = -g \cdot M_a \cdot CON4 \cdot CON9$$

$$CON21 = g \cdot M_a \cdot CON4 \cdot CON9/2$$

$$CON22 = -g \cdot M_d \cdot y_d$$

$$CON23 = g \cdot M_d \cdot y_d/2$$

$$CON24 = g \cdot M_d \cdot z_d$$

$$CON25 = -g \cdot M_b \cdot y_t$$

$$CON26 = g \cdot M_b \cdot y_t/2$$

$$CON27 = g \cdot M_b \cdot z_t$$

$$CON28 = -g \cdot M_b \cdot \eta_b \cdot CON9$$

$$CON29 = g \cdot M_b \cdot \eta_b \cdot CON9/2$$

$$CON30 = g \cdot M_b \cdot \zeta_b \cdot CON8$$

$$CON31 = -g \cdot M_b \cdot \zeta_b \cdot CON8/2$$

$$CON32 = g \cdot M_a \cdot CON8$$

$$CON33 = g \cdot M_a \cdot CON4 \cdot CON8$$

$$CON34 = M_b \cdot g \cdot \eta_b \cdot CON8$$

$$CON35 = M_b \cdot g \cdot \zeta_b \cdot CON9$$

$$CON36 = CON20 + CON22 + CON25 + CON28 + CON30$$

$$CON37 = CON21 + CON23 + CON26 + CON29 + CON31$$

$$CON38 = CON24 + CON27 + CON33 + CON34 + CON35 - k$$

$$CON39 = CON15 + CON36$$

$$CON40 = -M_a \cdot g \cdot CON9$$

With these definitions, Eqs. 8-206 and 8-207 may now be written as

$$\begin{aligned} CON1 \cdot \ddot{\eta} + CON2 \cdot \ddot{\phi} = {} & R(t) - B(t) \\ & + CON3 + CON10 \\ & + CON11 \cdot \dot{\phi}^2 \\ & + M_a \cdot \eta \cdot \dot{\phi}^2 \\ & + CON12 \cdot \dot{\phi}^2 \\ & + \phi \cdot CON40 \end{aligned} \tag{8-210}$$

$$\begin{aligned} [M_a(\eta + CON4)^2 + CON5]\ddot{\phi} = {} & \\ & CON13 \cdot \dot{\eta} \cdot \dot{\phi} \cdot \eta + CON14 \cdot \dot{\eta}\dot{\phi} \\ & + CON38 \cdot \phi + CON32 \cdot \eta\phi \\ & + CON39 - c\dot{\phi} + CON16 \cdot \dot{\phi}^2 \eta \\ & + CON17 \cdot \dot{\phi}^2 + B(t) \cdot CON7 \\ & + [R(t) + CON3] \cdot CON6 + CON18 \cdot \eta \\ & + CON19 \cdot \eta \cdot \phi^2 + CON37 \cdot \phi^2 \end{aligned} \tag{8-211}$$

Eqs. 8-210 and 8-211 can be put into the following form

$$\left.\begin{aligned} \nu_{11}\ddot{\eta} + \nu_{12}\ddot{\phi} &= \nu_{13} \\ \nu_{22}\ddot{\phi} &= \nu_{23} \end{aligned}\right\} \quad (8\text{-}212)$$

where

$\nu_{11} = \text{CON1}$

$\nu_{12} = \text{CON2}$

$$\begin{aligned} \nu_{13} = {} & R(t) - B(t) + \text{CON3} + \text{CON10} \\ & + \text{CON11}\cdot\phi^2 + M_a\eta\dot{\phi}^2 \\ & + \text{CON12}\cdot\dot{\phi}^2 + \text{CON40}\cdot\phi \end{aligned}$$

$\nu_{21} = 0$

$\nu_{22} = M_a(\eta + \text{CON4})^2 + \text{CON5}$

$$\begin{aligned} \nu_{23} = {} & \text{CON13}\cdot\dot{\eta}\dot{\phi}\eta + \text{CON14}\cdot\dot{\eta}\dot{\phi} \\ & + \text{CON38}\cdot\phi + \text{CON32}\cdot\eta\phi \\ & + \text{CON39} - \iota\dot{\phi} + \text{CON16}\cdot\dot{\phi}^2\eta \\ & + \text{CON17}\cdot\dot{\phi}^2 + B(t)\cdot\text{CON7} \\ & + [R(t) + \text{CON3}]\cdot\text{CON6} \\ & + \text{CON18}\cdot\eta + \text{CON19}\cdot\eta\cdot\phi^2 \\ & + \text{CON37}\cdot\phi^2 \end{aligned}$$

Eq. 8-212 can now be written as

$$\left.\begin{aligned} \ddot{\eta} &= (\nu_{13}\cdot\nu_{22} - \nu_{12}\nu_{23})/(\nu_{11}\cdot\nu_{22}) \\ \ddot{\phi} &= \nu_{23}/\nu_{22} . \end{aligned}\right\} (8\text{-}213)$$

By making the following definitions, Eqs. 8-213 can be put into first order form:

$$\left.\begin{aligned} x_1 &= \dot{\eta} \\ x_2 &= \eta \\ x_3 &= \dot{\phi} \\ x_4 &= \phi \end{aligned}\right\} \quad (8\text{-}214)$$

When this is accomplished, the following first-order equations yield the proper formulation that will be used in the steepest-descent scheme:

$$\left.\begin{aligned} \dot{x}_1 &= (\nu_{13}\cdot\nu_{22} - \nu_{12}\cdot\nu_{23}) \\ & \quad /(\nu_{11}\cdot\nu_{22}) \equiv f_1 \\ \dot{x}_2 &= x_1 \equiv f_2 \\ \dot{x}_3 &= \nu_{23}/\nu_{22} \equiv f_3 \\ \dot{x}_4 &= x_3 \equiv f_4 \end{aligned}\right\} \quad (8\text{-}215)$$

8-5.3 STEEPEST DESCENT FORMULATION

The optimal design problem can be stated as follows: Determine the design variable $R(t)$ in the interval $0 < t < t^3$ so as to

$$\text{minimize } J = x_4(t_2) \quad (8\text{-}216)$$

subject to the constraints

$$\left.\begin{aligned} \psi_1 &= x_2(t^1) - \eta_0 + \eta_{max} = 0 \\ \psi_2 &= x_2(t^3) - \eta_0 = 0 \\ \psi_3 &= x_1(t^3) - V_3 = 0 \\ \Omega^1 &= x_1(t^1) = 0 \\ \Omega^2 &= x_3(t^2) = 0 \\ \Omega^3 &= t^3 - c_T = 0 \end{aligned}\right\} \quad (8\text{-}217)$$

$$\phi = R(t) - R_{max} < 0 \tag{8-218}$$

and satisfying

$$\dot{x} = f \tag{8-219}$$

where the components of the vector f are given in Eq. 2-215 and initial conditions are

$$x_1(0) = x_3(0) = x_4(0) = 0,\ x_2(0) = \eta_0 .$$

The minimization problem stated here starts with an estimated design for $R(t)$, analyzes it, and then improves on the design. The steepest descent technique of par. 8-3 is used here to solve the design problem stated. The first step in implementing the computational algorithm of par. 8-3.3 is computation of auxiliary variables required for the algorithm.

8-5.3.1 DETERMINATION OF THE ADJOINT EQUATIONS

The adjoint equations are, from Eq. 8-93

$$\dot{\lambda} = -\left[\frac{\partial f}{\partial x}\right]^T \lambda,\ 0 < t < t^3,$$

where the vectors f and x are defined in Eqs. 8-215 and $\lambda = (\lambda_1, \lambda_2, \lambda_3, \lambda_4)^T$.

$$\frac{\partial f_1}{\partial x_i} = \left\{(\nu_{11}\nu_{22})\left[\nu_{13}\left(\frac{\partial \nu_{22}}{\partial x_i}\right) + \nu_{22}\left(\frac{\partial \nu_{13}}{\partial x_i}\right) - \nu_{12}\left(\frac{\partial \nu_{23}}{\partial x_i}\right) - \nu_{23}\left(\frac{\partial \nu_{12}}{\partial x_i}\right)\right] - (\nu_{13}\nu_{22} - \nu_{12}\nu_{23}) \times \left[\nu_{11}\left(\frac{\partial \nu_{22}}{\partial x_i}\right) + \nu_{22}\left(\frac{\partial \nu_{11}}{\partial x_i}\right)\right]\right\} / (\nu_{11}^2\nu_{22}^2) \quad i = 1,2,3,4$$

$$\frac{\partial f_1}{\partial x_1} = -\nu_{12}x_3(\mathrm{CON13}\cdot x_2 + \mathrm{CON14}) / (\nu_{11}\nu_{22})$$

$$\frac{\partial f_1}{\partial x_2} = \left\{\nu_{11}\nu_{22}[2M_a\nu_{13}(x_2 + \mathrm{CON4}) + M_a\nu_{22}x_3^2 - \nu_{12}(\mathrm{CON13}\cdot x_1x_3 + \mathrm{CON32}\cdot x_4 + \mathrm{CON16}\cdot x_3^2 + \mathrm{CON18} + \mathrm{CON19}\cdot x_4^2)] - (\nu_{13}\nu_{22} - \nu_{12}\nu_{23}) \times [2M_a\nu_{11}(x_2 + \mathrm{CON4})]\right\} / (\nu_{11}^2\nu_{22}^2)$$

$$\frac{\partial f_1}{\partial x_3} = [\nu_{22}(2M_a x_2x_3 + 2\cdot\mathrm{CON12}x_3) - \nu_{12}(\mathrm{CON13}\cdot x_1x_2 + \mathrm{CON14}\cdot x_1 - C + 2\cdot\mathrm{CON16}\cdot x_3x_2 + 2\cdot\mathrm{CON17}\cdot x_3)] / (\nu_{11}\nu_{22})$$

$$\frac{\partial f_1}{\partial x_4} = [\nu_{22}(2\cdot\mathrm{CON11}\cdot x_4 + \mathrm{CON40}) - \nu_{12}(\mathrm{CON38} + \mathrm{CON32}\cdot x_2 + 2\cdot\mathrm{CON19}\cdot x_2x_4 + 2\cdot\mathrm{CON37}\cdot x_4)] / (\nu_{11}\nu_{22})$$

$$\frac{\partial f_2}{\partial x_1} = 1,\ \frac{\partial f_2}{\partial x_2} = 0,\ \frac{\partial f_2}{\partial x_3} = 0,\ \frac{\partial f_2}{\partial x_4} = 0$$

$$\frac{\partial f_3}{\partial x_i} = \left[\nu_{22}\left(\frac{\partial \nu_{23}}{\partial x_i}\right) - \nu_{23}\left(\frac{\partial \nu_{22}}{\partial x_i}\right)\right] / \nu_{22}^2,$$

$i = 1,2,3,4$

$$\frac{\partial f_3}{\partial x_1} = x_3\,(\text{CON13}\cdot x_2 + \text{CON14})/\nu_{22}$$

$$\frac{\partial f_3}{\partial x_2} = [\nu_{22}\,(\text{CON13}\cdot x_1 x_3 + \text{CON32}\cdot x_4 + \text{CON16}\cdot x_3^2 + \text{CON18} + \text{CON19}\cdot x_4^2) - 2M_a \nu_{23}(x_2 + \text{CON4})]/\nu_{22}^2$$

$$\frac{\partial f_3}{\partial x_3} = (\text{CON13}\cdot x_1 x_2 + \text{CON14}\cdot x_1 - C + 2\cdot\text{CON16}\cdot x_3 x_2 + 2\,\text{CON17}\cdot x_3)/\nu_{22}$$

$$\frac{\partial f_3}{\partial x_4} = (\text{CON38} + \text{CON32}\cdot x_2 + 2\cdot\text{CON19}\cdot x_2 x_4 + 2\cdot\text{CON37}\cdot x_4)/\nu_{22}$$

$$\frac{\partial f_4}{\partial x_1} = 0,\ \frac{\partial f_4}{\partial x_2} = 0,\ \frac{\partial f_4}{\partial x_3} = 1,\ \frac{\partial f_4}{\partial x_4} = 0.$$

The adjoint equations now become

$$\dot{\lambda} = -\begin{bmatrix} \frac{\partial f_1}{\partial x_1} & 1 & \frac{\partial f_3}{\partial x_1} & 0 \\ \frac{\partial f_1}{\partial x_2} & 0 & \frac{\partial f_3}{\partial x_2} & 0 \\ \frac{\partial f_1}{\partial x_3} & 0 & \frac{\partial f_3}{\partial x_3} & 1 \\ \frac{\partial f_1}{\partial x_4} & 0 & \frac{\partial f_3}{\partial x_4} & 0 \end{bmatrix} \lambda \qquad (8\text{-}220)$$

where the partial derivatives are as computed in this paragraph.

8-5.3.2 DETERMINATION OF THE BOUNDARY CONDITIONS FOR THE ADJOINT EQUATIONS

Because of the intermediate constraint functions, we must evaluate λ at t^{2-} and t^{1-} to allow for any discontinuities which may occur across t^2 and t^1. Since the initial conditions for the adjoint equations are given at t^3, these equations are integrated backwards on the time interval shown in Fig. 8-5. Integration is carried out by integrating from t^3 to t^{2+}. Application of Eq. 8-116 provides new initial

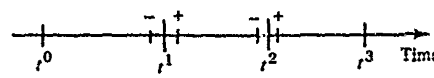


Figure 8-5. Recoil Time Interval

conditions at t^{2-}. Integration is then performed from t^{2-} to t^{1+}. Likewise, using new initial conditions at t^{1-}, integration is then performed to t^0.

It is the object of this paragraph to determine the initial conditions at t^3, t^{2-}, and t^{1-} for the four different integrations performed on the adjoint equations, i.e., for ψ_1, ψ_2, ψ_3, and J.

Since Ω^3 of Eq. 8-203 does not depend explicitly on x, $\partial\Omega^3/\partial x = 0$ and Eq. 8-114 reduces to

$$\lambda(t^3 - 0) = \frac{\partial g^T}{\partial x^3}.$$

Thus,

$$\left.\begin{aligned} \lambda^J(t^3) &= (0,0,0,0)^T \\ \lambda^{\psi_1}(t^3) &= (0,0,0,0)^T \\ \lambda^{\psi_2}(t^3) &= (0,1,0,0)^T \\ \lambda^{\psi_3}(t^3) &= (1,0,0,0)^T \end{aligned}\right\} \qquad (8\text{-}221)$$

Integration of Eq. 8-220 backward from t^3 to $(t^2 + 0)$ can now be effected, using the initial conditions of Eq. 8-221. Since Eq. 8-220 is homogeneous in λ and the initial conditions on λ^J and λ^{ψ_1} are zero, it is clear that $\lambda^J(t^2 + 0) = \lambda^{\psi_1}(t^2 + 0) = 0$. While $\lambda^{\psi_2}(t^2 + 0)$ and $\lambda^{\psi_3}(t^2 + 0)$ will not be zero, they may be treated as known. There is now adequate information to use Eq. 8-116 to determine $\lambda(t^2 - 0)$ for all four adjoint variables.

In order to find $\lambda(t^2 - 0)$ in Eq. 8-116, choose the minus sign alternative throughout and obtain

$$\lambda(t^2 - 0) = \lambda(t^2 + 0) + \left(\frac{\partial g^T}{\partial x^2}\right) - \left(\frac{1}{\dot{\Omega}^2(t^2 - 0)}\right)\Big\{\dot{g}(t^2 - 0) + \lambda^T(t^2 + 0)\,[f(t^2 - 0) - f(t^2 + 0)]\Big\}\left(\frac{\partial \Omega^{2\,T}}{\partial x^2}\right) \tag{8-222}$$

Using Eqs. 8-107 and 8-109 to determine $\dot{g}(t^2 - 0)$ and $\dot{\Omega}(t^2 - 0)$, one obtains at $t^2 - 0$

$$\dot{g}^J(t^2 - 0) = f_4(t^2 - 0)$$

$$\dot{g}_1(t^2 - 0) = 0$$

$$\dot{g}_2(t^2 - 0) = 0$$

$$\dot{g}_3(t^2 - 0) = 0$$

and

$$\dot{\Omega}_2(t^2 - 0) = f_3(t^2 - 0).$$

Thus, Eq. 8-222 yields

$$\left.\begin{aligned}
&\lambda^J(t^2 - 0) = \begin{bmatrix}0\\0\\0\\1\end{bmatrix} - \left[\frac{f_4(t^2 - 0)}{f_3(t^2 - 0)}\right]\begin{bmatrix}0\\0\\1\\0\end{bmatrix}\\
&\lambda^{\psi_1}(t^2 - 0) = \begin{bmatrix}0\\0\\0\\0\end{bmatrix}\\
&\lambda^{\psi_2}(t^2 - 0) = \lambda^{\psi_2}(t^2 + 0)\\
&\qquad \left\{\frac{\lambda^{\psi_2 T}(t^2 + 0)[f(t^2 - 0) - f(t^2 + 0)]}{f_3(t^2 - 0)}\right\}\begin{bmatrix}0\\0\\1\\0\end{bmatrix}\\
&\text{and}\\
&\lambda^{\psi_3}(t^2 - 0) = \lambda^{\psi_3}(t^2 + 0)\\
&\qquad - \left\{\frac{\lambda^{\psi_3 T}(t^2 + 0)[f(t^2 - 0) - f(t^2 + 0)]}{f_3(t^2 - 0)}\right\}\begin{bmatrix}1\\0\\1\\0\end{bmatrix}
\end{aligned}\right\} \tag{8-223}$$

It may be noted that $\lambda^{\psi_1}(t^1 + 0) = 0$ will result from integration of the homogeneous Eq. 8-220.

Finally, at t^1, Eqs. 8-107 and 8-109 yield

$$\dot{g}^J(t^1 - 0) = 0$$

$$\dot{g}^{\psi_1}(t^1 - 0) = f_2(t^1 - 0)$$

$$\dot{g}^{\psi_2}(t^1 - 0) = 0$$

$$\dot{g}^{\psi_3}(t^1 - 0) = 0$$

and

$$\dot{\Omega}^1(t^1 - 0) = f_1(t^1 - 0)$$

Replacing t^2 with t^1, and x^2 with x^1 in Eq. 8-222 provides the proper jump conditions at t^1. They are

$$
\left.\begin{aligned}
\lambda^J(t^1-0) &= \lambda^J(t^1+0) - \left\{\frac{\lambda^{J^T}(t^1+0)[f(t^1-0)-f(t^1+0)]}{f_1(t^1-0)}\right\}\begin{bmatrix}1\\0\\0\\0\end{bmatrix} \\
\lambda^{\psi_1}(t^1-0) &= \begin{bmatrix}0\\1\\0\\0\end{bmatrix} - \left[\frac{2}{f_1(t^1-0)}\right]\begin{bmatrix}1\\0\\0\\0\end{bmatrix} \\
\lambda^{\psi_2}(t^1-0) &= \lambda^{\psi_2}(t^1+0) - \left\{\frac{\lambda^{\psi_2^T}(t^1+0)[f(t^1-0)-f(t^1+0)]}{f_1(t^1-0)}\right\}\begin{bmatrix}1\\0\\0\\0\end{bmatrix} \\
\lambda^{\psi_3}(t^1-0) &= \lambda^{\psi_3}(t^1+0) - \left\{\frac{\lambda^{\psi_3^T}(t^1+0)[f(t^1-0)-f(t^1+0)]}{f_1(t^1-0)}\right\}\begin{bmatrix}1\\0\\0\\0\end{bmatrix}
\end{aligned}\right\} \quad (8\text{-}224)
$$

Eq. 8-220 may now be integrated from t^3 to t^0, with jumps at t^2 and t^1 defined by Eqs. 8-223 and 8-224. This completes computation required by Step 5 of the algorithm of par. 8-3.3.

8-5.3.3 COMPUTATION OF DESIGN IMPROVEMENTS

The remaining steps of the computational algorithm of par. 8-3.3 require only routine calculation. Some of the key formulas are highlighted here to illustrate use of the algorithm. In Step 6, the following calculations are effected:

$$\Lambda^J(t) = \frac{\partial f^T}{\partial R}\lambda^J(t)$$

$$\Lambda^{\psi_i}(t) = \frac{\partial f^T}{\partial R}\lambda^{\psi_i}(t) \qquad i = 1,2,3,$$

$$\Lambda^\phi(t) = \begin{cases} W_R^{-1}(t), & \text{if } \phi > 0 \\ 0\ , & \text{if } \phi < 0. \end{cases}$$

where $W_R(t)$ is a weighting factor, set equal to one in this example.

With these factors defined, one must choose the magnitude of constraint error correction to be used, in Step 7. In the current problem, reasonable design estimates led to small errors, so $a = c = 1$ was chosen.

For Step 3, only the following routine numerical integrations are required:

$$M_{\psi J} = \int_{t^0}^{t^3} \Lambda^{\psi^T} d(t)\, \Lambda^J\, dt$$

where

$$d(t) = \begin{cases} 1, & \text{if } \phi < 0 \\ 0, & \text{if } \phi > 0 \end{cases}$$

$$M_{\psi\psi} = \int_{t^0}^{t^3} \Lambda^{\psi^T} d(t)\, \Lambda^\psi\, dt$$

$$M_{\psi\phi} = \int_{t^0}^{t^1} \Lambda^{\psi^T} \bar{\phi}\, dt.$$

Eq. 8-80 was used to find γ_0 so that a ten percent reduction in cost function is sought. From Eq. 8-43 (since $a = c = 1$)

$$\gamma = M_{\psi\psi}^{-1}\,[2\gamma_0(\bar{\psi} - M_{\psi\phi}) - M_{\psi J}]$$

and from Eq. 8-35

$$\mu(t) = -\Lambda^\phi(t)^{-1}\,[(1 - d(t))(\Lambda^J + \Lambda^\psi\gamma) - 2\gamma_0\phi].$$

At any points where $\mu(t) < 0$, delete this point from the domain of $\bar{\phi}(t)$ and return to Step 6.

Finally, the improved design is provided by Step 11 of the algorithm, where entries are computed directly from Eqs. 8-44 and 8-45. While these formulas are a bit messy, they are routinely programmed matrix computations which are no real challenge to the computer.

8-5.4 RESULTS AND CONCLUSIONS

Fig. 8-6 represents the optimal recoil force to minimize hop at 75-deg quadrant elevation

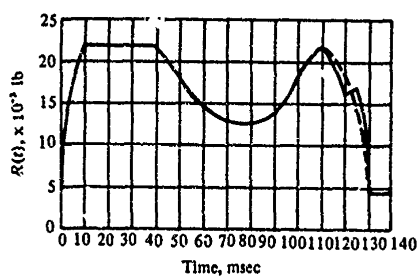


Figure 8-6. Optimal Rod Force

with the following constraints:

$$\left.\begin{array}{l} R(t) < 22000 \text{ lb} \\ \text{recoil length} = 28 \text{ in.} \end{array}\right\} \quad (8\text{-}225)$$

The resulting hop for this case is 1.53 in., i.e., the tires leave the ground 1.53 in. for a 115% maximum rated pressure breech force. If the constraints were relaxed, such that

$$\left.\begin{array}{l} R(t) < 23500 \text{ lb} \\ \text{recoil length} = 29 \text{ in.} \end{array}\right\} \quad (8\text{-}226)$$

the resulting minimal hop is 0.88 in.

The acceleration of the recoiling parts during the first portion of counterrecoil is an important factor in reducing the hop; i.e., the faster the recoiling parts accelerate during this period, the greater the reduction in hop. As one would expect, an increase in allowable recoil length also reduces hop significantly. An increase in the maximum rod force will also reduce hop, for example, if the recoil force is allowed to obtain the value 24160 lb in constraint set of Eq. 8-225, the hop can be reduced an additional 0.32 in. Fig. 8-7 shows

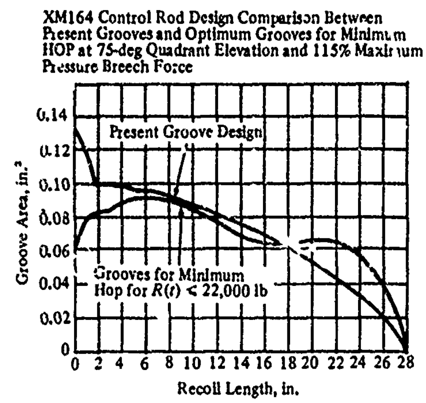


Figure 8-7. Optimal Control Rod Design

a possible variable orifice area design for short recoil. The orifice areas were obtained from the recoil force in Fig. 8-6. The resulting force levels from the new groove design are indicated by the dotted lines from 0.110 sec to 0.13 sec, Fig. 8-6. The recoil force is the same as the optimal shaped force curve from 0 to 0.110 sec. The increase in hop is approximately 0.1 in. The recoil length changed a very small amount.

An interesting side point is that of the speed of convergence. The nominal design variable $R(t)$ used for the first iteration was such that at the end of counterrecoil the recoiling parts were 250 in. away from the latch position and the required final velocity of 6 in./sec, was 96 in./sec. In approximately 14 iterations, convergence was obtained which seems to be very fast if one considers the complexity of the equations involved.

Results from firing tests show a significant reduction (50% or more) in hop can be achieved simply by increasing the tire pressure. Because tire performance information is not presently available, it was assumed throughout this analysis that the spring rate of the tires was constant (linear spring). Therefore, it is not known what results would be obtained under a nonlinear spring model. Tire manufacturers are investigating methods to optimize tire characteristics for the final configuration in the tire itself. In order to obtain optimum weapon performance for flexible mount systems, such information as tire performance could be incorporated into the mathematical model and perhaps tire characteristics could also be optimized in the environment for which they are being used.

The technique used here has the capability to optimize many design parameters simultaneously. If there exist other sensitive parameters, consideration should be given to optimize them along with the design variable $R(t)$. This study clearly indicates that weapon performance can be improved by using methods of optimal design.

REFERENCES

1. R. Courant and D. Hilbert, *Methods of Mathematical Physics* Vol. 2, Interscience, New York, 1962.

2. C. Lanczos, *Linear Differential Operators*, Van Nostrand, London, 1961.

3. M. R. Hestenes, *Calculus of Variations and Optimal Control Theory*, John Wiley & Sons, New York, 1966.

4. H. J. Kelley, "Methods of Gradients", in G. Leitmann, Ed., *Optimization Techniques*, Academic Press, New York, 1962.

5. W. F. Denham, *Steepest-Ascent Solution of Optimal Programming Problems*, Report No. BR-2393, Raytheon Co., Bedford, Mass., April, 1963.

6. R. E. Kopp, and H. G. Moyer, "Trajectory Optimization Techniques", in C. T. Leondes, Ed., *Advances in Control Systems*, Vol. 4, Academic Press, New York, 1966.

7. D. S. Hague, "Solution of Multiple-Arc Problems by the Steepest-Descent Method", in A. Lavi and T. P. Vogl, Eds., *Recent Advances in Optimization Techniques*, Wiley, New York, 1966.

8. B. D. Tapley, and J. M. Lewallen, "Comparison of Several Numerical Optimization Methods", *J. Opt. Theory and Appl.*, Vol 1, 1967, pp. 1-32.

9. A. P. Sage, *Optimum Systems Control*, Prentice-Hall, Englewood Cliffs, New Jersey, 1968.

10. E. J. Haug Jr., R. S. Newell, and T. D. Streeter, *Optimal Design of Elastic Structural Elements*, Tech. Rept. SY-R1-69,

Systems Analysis Directorate, Headquarters, US Army Weapons Command, Rock Island, Illinois, 1969.

11. E. B. Lee and L. Markus, *Foundations of Optimal Control Theory*, Wiley, New York, 1967.

12. P. K. Wang "Control of Distributed Parameter Systems", in C. T. Leondes, Ed., *Advances in Control Systems*, Vol. 1, Academic Press, New York, 1964.

13. W. L. Brogan, *Optimal Control Theory Applied to Systems Described by Partial Differential Equations*, Report No. 65-29, Department of Engineering, UCLA, Los Angeles, 1965.

14. L. A. Liusternik and V. J. Sobolev, *Elements of Functional Analysis*, Ungar New York, 1961.

15. T. D. Streeter, *Optimal Weapon Stability By a Steepest-Descent Method*, Tech. Rept. SY-R2-69, Systems Analysis Directorate, U.S. Army Weapons Command, Rock Island, Ill, Aug 1969.

16. J. W. Frantz and R. H. Coberly, *Design of Control Rod for XM44 Recoil Mechanism (XM164 Howitzer) Using Coupled Rotational and Translational Equations of Motion*, Technical Notes, U.S. Army Weapons Command, Rock Island, Illinois, August 1968.

CHAPTER 9

APPLICATION OF STEEPEST DESCENT METHODS TO OPTIMAL STRUCTURAL DESIGN

9-1 INTRODUCTION

The same general class of optimal structural design problems considered in Chapter 5 is treated in this chapter by using the theoretical results and computational algorithms developed in Chapter 8. For a discussion of the essential elements of the design problem to be considered, the reader is referred to par. 5-2.

The basic difference between problems treated in this chapter and those treated in Chapter 5 is in the nature of the design variables and the associated state variables that describe the response of the structure. In this chapter, distribution of the material along members of the structure will be continuous as opposed to discrete, as treated in Chapter 5. Consequently, continuous variation of stress, displacement, and eigenfunctions along members of the structure will need to be determined. In this sense, the present problem is infinite dimensional whereas the problem treated in Chapter 5 was finite dimensional. The optimal structural design problem, therefore, will involve boundary value problems as opposed to algebra problems.

The continuous design problem treated in this chapter will have more features to account for than did the discrete problem. For example, there will be differential equations, boundary conditions, pointwise inequality constraints, and functional constraints. Further, nonclassical analysis problems may arise which require special techniques for each particular problem. For this reason, it is more difficult to give a general formulation of the problem into which every structural design problem will fit.

Examples of difficulties that may occur in particular problems include the dependence of boundary conditions on design parameters and design variables, inequality constraints for stress which involve state and design variables that must be transformed to functional constraints, and interrelationships between eigenvalues and state variables at particular points in the structure. Each of these peculiar features will be treated as it arises in a particular problem. This requires that the designer who is using continuous optimization methods for structural design must understand the origin of the methods well enough so that he can alter the computational algorithms as required to fit his particular problems.

While the previous discussion might indicate that one encounters only difficulties when using continuous methods as opposed to discrete methods, there appears to be a potential for more efficient computational methods in the infinite dimensional problem than in the finite dimensional formulation. Further, it is clear that the infinite dimensional formulation can yield a true optimum while the discrete formulation of the problem will generally yield only an approximate optimum. In the following paragraph, a relatively general formulation of the infinite dimensional optimal structural design problem is given, and the computational algorithm based on theory of Chapter 8 is developed. In following paragraphs, this theory and computational algorithm are applied to example problems. Alterations in the general theory are made as they are required in the solution of individual problems.

9-2 STEEPEST DESCENT METHOD FOR OPTIMAL STRUCTURAL DESIGN

In the structural design problems treated here, the design is to be specified by a vector design function $u(x) = [u_1(x), ..., u_m(x)]^T$ and a vector design parameter $b = [b_1, ..., b_p]^T$, where the independent variable x may be a real variable, a two dimensional variable $x = [x_1, x_2]^T$, or a three dimensional variable $x = [x_1, x_2, x_3]^T$, depending on whether material is to be distributed over a line, a surface, or a volume. In addition to the design variables $u(x)$ and b, there will be state variables $z(x) = [z_1(x), ..., z_n(x)]^T$ representing stress and displacement under load and $y(x) = [y_1(x), ..., y_q(x)]^T$ representing mode shapes for vibration or buckling.

For the purposes of convenience in notation and generality, the system equations will be written in operator notation similar to that used in par. 8-4. Only linear behavior will be considered so that stress and deflection are determined by the linear boundary-value problem

$$L(u,b)z = Q(u,b), \qquad x \in \Omega \tag{9-1}$$

and

$$Bz = q, \qquad x \in \Gamma \tag{9-2}$$

In this notation, Ω is the region over which the material of the structure is distributed and Γ is the boundary of that region. L and B are differential operators on Ω and Γ, respectively. The functions Q and q are generally related to loads.

To better fix the idea of operators, consider the simply supported beam of Fig. 9-1.

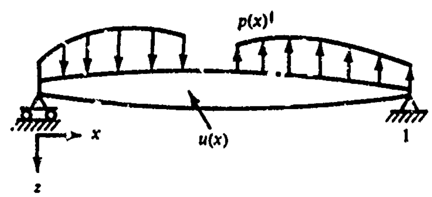


Figure 9-1. Simply Supported Beam

The boundary value problem in this case is simply

$$L(u)z = EI(u)\frac{d^2z}{dx^2} = -M(x) \equiv Q \tag{9-3}$$

and

$$Bz \equiv \begin{bmatrix} z(0) \\ \\ z(1) \end{bmatrix} = \begin{bmatrix} 0 \\ \\ 0 \end{bmatrix} \equiv q. \tag{9-4}$$

Here, $u(x)$ is a variable which uniquely specified the beam cross section and determines the design, E is Young's modulus, $z(x)$ is deflection, and $M(x)$ is the bending moment that is computed from the distributed load $p(x)$.

In this example, Ω is just the interval $(0,1)$ and Γ consists of the two endpoints $x = 0$ and $x = 1$. The advantage in the notation of Eqs. 9-1 and 9-2 is that it is convenient and at the same time applied to a large class of problems. For an example of a problem in which Ω is a subset of two dimensional space, see par. 9-7.

In addition to response due to static load, it is necessary to treat the response of a structure to dynamic loads. One important characteristic of a structure, which is classified as dynamic response, is natural frequency. Another response, which must be treated, is buckling. Both buckling loads and natural frequencies are determined, in the present class of problems, by linear eigenvalue problems. Again using operator notation, these problems may be written in the form

$$K(u,b)y = \zeta M(u,b)y, \quad x \in \Omega \tag{9-5}$$

and

$$Cy = 0, \qquad x \in \Gamma \tag{9-6}$$

where ζ is the eigenvalue and $y(x)$ the associated eigenfunction or mode shape. The operators K and M generally relate to stiffness and mass, respectively. In conservative problems they will be symmetric (Ref. 1) (formally self-adjoint) but in nonconservative systems, they will not be symmetric. The more general unsymmetric case is treated here.

The optimal design problem is that of minimizing

$$J = f_0(b,\zeta) + \iint_{\Omega} f_1(x,z,u,b)\,d\Omega \tag{9-7}$$

subject to the pointwise constraints

$$\phi_i(x,u) < 0, \quad x \in \Omega, \quad i = 1, \ldots, r \tag{9-8}$$

and the functional constraints

$$\psi_j = e_j(b,\zeta) + \iint_{\Omega} g_j(x,z,u,b)\,d\Omega < 0,$$

$$j = 1, \ldots, s \tag{9-9}$$

Constraints of the form $\eta(x,z,u,b) < 0$ for $x \in \Omega$ will have to be reduced to functional constraints as in Eq. 8-6.

Beginning with an engineering estimate of the design variables $u(x)$ and b, Eqs. 9-3 through 9-6 may be solved for $z(x)$, $y(x)$, and ζ. A perturbation, $(\delta u, \delta b)$, in the design leads to perturbations in the cost and constraint functions

$$\delta J = \frac{\partial f_0}{\partial b}\delta b + \frac{\partial f_0}{\partial \zeta}\delta\zeta + \iint_{\Omega}\left(\frac{\partial f_1}{\partial z}\delta z + \frac{\partial f_1}{\partial u}\delta u + \frac{\partial f_1}{\partial b}\delta b\right)d\Omega \tag{9-10}$$

$$\delta\phi_i = \frac{\partial \phi_i}{\partial u}\delta u, \quad x \in \Omega, \quad i = 1, \ldots, r \tag{9-11}$$

$$\delta\psi_j = \frac{\partial e_j}{\partial b}\delta b + \frac{\partial e_j}{\partial \zeta}\delta\zeta + \iint_{\Omega}\left(\frac{\partial g_j}{\partial z}\delta z + \frac{\partial g_j}{\partial u}\delta u + \frac{\partial g_j}{\partial b}\delta b\right)d\Omega,$$

$$j = 1, \ldots, s. \tag{9-12}$$

The object, as in preceding work, is to eliminate explicit dependencies on δz and $\delta\zeta$. First, the perturbation equation for δz is

$$\left.\begin{aligned} &L(u,b)\delta z + \frac{\partial}{\partial u}[L(u,b)z]\delta u \\ &\quad + \frac{\partial}{\partial b}[L(u,b)z]\delta b = \frac{\partial Q}{\partial u}\delta u \\ &\quad + \frac{\partial Q}{\partial b}\delta b, \quad x \in \Omega \\ &B\delta z = 0, \quad x \text{ on } \Gamma. \end{aligned}\right\} \tag{9-13}$$

In certain problems, the boundary conditions may depend on the design parameter

b. In that case, the perturbed boundary condition is

$$B\delta z + \frac{\partial}{\partial b}[B(b)z]\delta b = 0$$

instead of Eq. 9-13. In this case, methods of par. 8-4 must be applied to particular problems.

To eliminate δz, integration of $\int\int_\Omega \lambda^T L\delta z\, d\Omega$ by parts yields operators L^* in Ω and B^* on Γ such that

$$\int\int_\Omega [\lambda^T L\delta z - \delta z^T L^*\lambda]d\Omega = 0 \qquad (9\text{-}14)$$

for all δz and λ satisfying $B\delta z = 0$ and $B^*\lambda = 0$ on Γ.

Solving

$$\left.\begin{aligned} L^*\lambda^J &= \frac{\partial f_1}{\partial z}^T, \quad x\in\Omega \\ B^*\lambda^J &= 0, \quad x\in\Gamma \end{aligned}\right\} \qquad (9\text{-}15)$$

and

$$\left.\begin{aligned} L^*\lambda^{\psi_j} &= \frac{\partial g_j}{\partial z}^T, \quad x\in\Omega \\ B^*\lambda^{\psi_j} &= 0, \quad x\in\Gamma \end{aligned}\right\} \qquad (9\text{-}16)$$

and substituting into Eqs. 9-10 and 9-12, using Eq. 9-13, one obtains

$$\delta J = \frac{\partial f_0}{\partial \zeta}\delta\zeta + \int\int_\Omega \left[\lambda^{J^T}\left(\frac{\partial Q}{\partial u} - \frac{\partial}{\partial u}[L(u,b)z]\right) + \frac{\partial f_1}{\partial u}\right]\delta u d\Omega + \left\{\frac{\partial f_0}{\partial b} + \int\int_\Omega\left[\frac{\partial f_1}{\partial b} + \lambda^{J^T}\left(\frac{\partial Q}{\partial b} - \frac{\partial}{\partial b}[L(u,b)z]\right)\right]d\Omega\right\}\delta b \qquad (9\text{-}17)$$

and

$$\delta\psi_j = \frac{\partial e_j}{\partial \zeta}\delta\zeta + \int\int_\Omega\left[\lambda^{\psi_j^T}\times\left(\frac{\partial Q}{\partial u} - \frac{\partial}{\partial u}[L(u,b)z]\right) + \frac{\partial g_j}{\partial u}\right]\delta u d\Omega + \left\{\frac{\partial e_j}{\partial b} + \int\int_\Omega\left[\frac{\partial g_j}{\partial b} + \lambda^{\psi_j^T}\left(\frac{\partial Q}{\partial b} - \frac{\partial}{\partial b}[L(u,b)z]\right)\right]d\Omega\right\}\delta b \qquad (9\text{-}18)$$

It remains only to eliminate explicit dependence on $\delta\zeta$. Under very restrictive hypotheses, Kato (Ref. 2) has obtained a relationship among $\delta\zeta$, δb, and δu. This relationship is derived here formally. It is assumed that $\delta\zeta$ and δy depend continuously on δb and δu and further that the following perturbation formula holds:

$$K(u,b)\delta y + \frac{\partial}{\partial u}[K(u,b)y]\delta u + \frac{\partial}{\partial b}[K(u,b)y]\delta b = \delta\zeta M(u,b)y + \zeta\frac{\partial}{\partial u}[M(u,b)y]\delta u + \zeta\frac{\partial}{\partial b}[M(u,b)y]\delta b + \zeta M(u,b)\delta y \qquad (9\text{-}19)$$

Just as in Eq. 9-14, integration by parts may be used to obtain the operators $K^* - \zeta M^*$ and C^* adjoint to $K - \zeta M$ and C. These operators are defined by the relation

$$\iint_\Omega w^T(Ky - \zeta My)d\Omega = \iint_\Omega \bar{y}^T(K^* w - \zeta M w)d\Omega$$

for all w and y satisfying $C^* w = 0$ and $Cy = 0$ on Γ. Let $\bar{y}$ satisfy

$$\left.\begin{aligned} K^*\bar{y} &= \zeta M^*\bar{y}, \quad x \in \Omega \\ C^*\bar{y} &= 0, \quad x \in \Gamma \end{aligned}\right\} \quad (9\text{-}20)$$

Premultiplying Eq. 9-19 by $\bar{y}^T$, integrating, and rearranging terms yields

$$\left\{\iint_\Omega \bar{y}^T My\, d\Omega\right\}\delta\zeta = \iint_\Omega [\bar{y}^T K\delta y - \zeta\bar{y}^T M\delta y]\, d\Omega + \iint_\Omega \left\{\bar{y}^T \frac{\partial}{\partial u}[K(u,b)y] - \zeta\bar{y}^T \frac{\partial}{\partial u}[M(u,b)y]\right\}\delta u\, d\Omega + \left\{\iint_\Omega \left(\bar{y}^T \frac{\partial}{\partial b}[K(u,b)y] - \zeta\bar{y}^T \frac{\partial}{\partial b}[M(u,b)y]\right)d\Omega\right\}\delta b \quad (9\text{-}21)$$

The first term on the right is

$$\iint_\Omega [\bar{y}^T K\delta y - \zeta\bar{y}^T M\delta y]\, d\Omega = \iint_\Omega [\delta y^T(K^*\bar{y} - \zeta M^*\bar{y})]\, d\Omega = 0$$

by definition of the adjoint operator. It must be emphasized that Eq. 9-21 is obtained by strictly formal calculations. There are deep mathematical questions concerning the general validity of this result. For a treatment of the subject of perturbation theory in linear operators, the reader is referred to Ref. 2.

An expression for $\delta\zeta$ in terms of δu and δb may not be substituted into Eqs. 9-17 and 9-18. Defining

$$\ell^J = \frac{\partial f_0^T}{\partial b} + \iint_\Omega \left\{\frac{\partial f_1^T}{\partial b} + \left[\frac{\partial Q^T}{\partial b} - \left(\frac{\partial}{\partial b}[L(u,b)z]\right)^T\right]\lambda^J + \left[\frac{\partial f_0}{\partial \zeta}\Big/\iint_\Omega \bar{y}^T My d\Omega\right] \times \left[\left(\frac{\partial}{\partial b}[K(u,b)y]\right)^T \bar{y} - \zeta\left(\frac{\partial}{\partial b}[M(u,b)y]\right)^T \bar{y}\right]\right\}d\Omega \quad (9\text{-}22)$$

$$\Lambda^J = \frac{\partial f_1^T}{\partial u} + \left[\frac{\partial Q^T}{\partial u} - \left(\frac{\partial}{\partial u}[L(u,b)z]\right)^T\right]\lambda^J + \left[\frac{\partial f_0}{\partial \zeta}\Big/\iint_\Omega \bar{y}^T My d\Omega\right] \times \left[\frac{\partial}{\partial u}\left([K(u,b)y]\right)^T \bar{y} - \zeta\left(\frac{\partial}{\partial u}[M(u,b)y]\right)^T \bar{y}\right] \quad (9\text{-}23)$$

$$\ell^{\psi_j} = \frac{\partial e_j^T}{\partial b} + \iint_\Omega \left\{ \frac{\partial g_j^T}{\partial b} + \left[\frac{\partial Q^T}{\partial b} - \left(\frac{\partial}{\partial b} [L(u,b)z] \right)^T \right] \lambda^{\psi_j} + \left[\frac{\partial e_j}{\partial \xi} \Big/ \iint_\Omega \bar{y}^T M y d\Omega \right] \times \left[\frac{\partial}{\partial b} \left([K(u,b)y] \right)^T \bar{y} - \xi \left(\frac{\partial}{\partial b} [M(u,b)y] \right)^T \bar{y} \right] \right\} d\Omega \tag{9-24}$$

and

$$\Lambda^{\psi_j} = \frac{\partial g_j^T}{\partial u} \left[\frac{\partial Q^T}{\partial u} - \left(\frac{\partial}{\partial u} [L(u,b)z] \right)^T \right] \lambda^{\psi_j} + \left[\frac{\partial e_j}{\partial \xi} \Big/ \iint_\Omega \bar{y}^T M y d\Omega \right] \times \left[\frac{\partial}{\partial u} \left([K(u,b)y] \right)^T \bar{y} - \xi \left(\frac{\partial}{\partial u} [M(u,b)y] \right)^T \bar{y} \right] \tag{9-25}$$

Eqs. 9-17 and 9-18 become

$$\delta J = \ell^{J^T} \delta b + \iint_\Omega \Lambda^{J^T} \delta u \; d\Omega \tag{9-26}$$

and

$$\delta \psi_j = \ell^{\psi_j^T} \delta b + \iint_\Omega \Lambda^{\psi_j} \delta u \; d\Omega . \tag{9-27}$$

Define vector constraint functions $\bar{\phi}(x)$ containing those constraint functions $\phi_i(x) > 0$ and $\bar{\psi}$ containing constraint functionals $\psi_j > 0$. Define

$$\Lambda^{\bar{\phi}}(x) = \left[\frac{\partial \phi_i^T}{\partial u} \mid \text{for all } i \text{ with } \phi_i(x) > 0 \right] \tag{9-28}$$

$$\ell^{\bar{\psi}} = \left[\ell^{\psi_j} \mid \text{for all } j \text{ with } \psi_j > 0 \right] \tag{9-29}$$

and

$$\Lambda^{\bar{\psi}}(x) = \left[\Lambda^{\psi_j} \mid \text{for all } j \text{ with } \psi_j > 0 \right] . \tag{9-30}$$

Defining $\Delta\bar{\phi}(x)$ and $\Delta\bar{\psi}$ to be desired reduction in constraint error, the linearized problem is to choose $\delta u(x)$ and δb to minimize

$$\delta J = \ell^{J^T} \delta b + \iint_\Omega \Lambda^{J^T} \delta u \; d\Omega \tag{9-31}$$

subject to

$$\Lambda^{\bar{\phi}^T} \delta u - \Delta\bar{\phi} < 0, \; \text{for } x \in \Omega \tag{9-32}$$

$$\ell^{\bar{\psi}^T} \delta b + \iint_\Omega \Lambda^{\bar{\psi}^T} \delta u \; d\Omega - \Delta\bar{\psi} < 0 \tag{9-33}$$

and

$$\delta b^T W_b \delta b + \iint_\Omega \delta u^T W_u(x) \delta u \; d\Omega - \xi^2 < 0 \tag{9-34}$$

The weighting matrices W_b and $W_u(x)$ are chosen positive definite and ξ is small. This optimization problem for δu and δb coincides with the problems of pars. 8-2 and 8-3 for ordinary differential equations and with the problem of par. 8-4 for partial differential equations.

Determination of a solution proceeds exactly as in Chapter 8 so only the resulting computational algorithm is given here.

Steepest Descent Algorithm for Optimal Structural Design

Step 1. Make an engineering estimate of the solution $u^{(0)}(x)$ and $b^{(0)}$.

Step 2. For $j = 0,1, ...,$ solve Eqs. 9-1, 9-2, 9-5, and 9-6 for $z^{(j)}$, $y^{(j)}$, and $\zeta^{(j)}$.

Step 3. Check the constraints of Eqs. 9-8 and 9-9 and form $\bar{\phi}$ and $\bar{\psi}$ containing the constraints not strictly satisfied.

Step 4. Solve the boundary value problems, Eqs. 9-15 and 9-16, for λ^J and $\lambda^{\psi j}$ corresponding to $\psi_j > 0$. Solve the eigenvalue problem, Eq. 9-20, for $\bar{y}$.

Step 5. Choose the corrections in constraint errors $\Delta\bar{\phi}$ and $\Delta\bar{\psi}$.

Step 6. Evaluate ℓ^J, Λ^J, $\ell^{\psi j}$, and $\Lambda^{\psi j}$ in Eqs. 9-22 through 9-25 and compute

$$M_{\phi\phi}(x) = \frac{\partial\bar{\phi}}{\partial u} W_u^{-1} \frac{\partial\bar{\phi}^T}{\partial u}$$

$$M_{\psi J} = \ell^{\bar{\psi}^T} W_b^{-1}\ell^J + \iint_\Omega \Lambda^{\bar{\psi}^T} W_u^{-1} \times \left(I - \frac{\partial\bar{\phi}^T}{\partial u} M_{\phi\phi}^{-1} \frac{\partial\bar{\phi}}{\partial u} W_u^{-1}\right)\Lambda^J \, d\Omega$$

$$M_{\psi\psi} = \ell^{\bar{\psi}^T} W_b^{-1}\ell^{\bar{\psi}} + \iint_\Omega \Lambda^{\bar{\psi}^T} W_u^{-1} \times \left(I - \frac{\partial\bar{\phi}^T}{\partial u} M_{\phi\phi}^{-1} \frac{\partial\bar{\phi}}{\partial u} W_u^{-1}\right)\Lambda^{\bar{\psi}} \, d\Omega$$

and

$$M_{\psi\phi} = \iint_\Omega \Lambda^{\bar{\psi}^T} W_u^{-1} \frac{\partial\bar{\phi}^T}{\partial u} M_{\phi\phi}^{-1} \Delta\bar{\phi} \, d\Omega$$

Step 7. Choose stepsize $\gamma_0 > 0$ and evaluate

$$\gamma = -M_{\psi\psi}^{-1}\left[2\gamma_0(\Delta\bar{\psi} + M_{\psi\phi}) + M_{\psi J}\right]$$

and

$$\mu(x) = -M_{\phi\phi}^{-1}\left[2\gamma_0\,\Delta\bar{\phi} + \frac{\partial\bar{\phi}}{\partial u} W_u^{-1}(\Lambda^J + \Lambda^{\bar{\psi}}\gamma)\right].$$

If any components of γ or $\mu(x)$ are negative, delete the corresponding components of $\bar{\psi}$ and $\bar{\phi}$, respectively, and return to Step 5. Otherwise, continue.

Step 8. Compute

$$\delta u^1(x) = W_u^{-1}\left(I - \frac{\partial\bar{\phi}^T}{\partial u} M_{\phi\phi}^{-1} \frac{\partial\bar{\phi}}{\partial u} W_u^{-1}\right) \times (\Lambda^J - \Lambda^{\bar{\psi}} M_{\psi\psi}^{-1} M_{\psi J})$$

$$\delta u^2(x) = W_u^{-1}\left(I - \frac{\partial \bar{\phi}^T}{\partial u} M_{\phi\phi}^{-1} \frac{\partial \bar{\phi}}{\partial u} W_u^{-1}\right) \times [\Lambda^{\bar{\psi}} M_{\psi\psi}^{-1} (\Delta\bar{\psi} + M_{\psi\phi})]$$

$$\delta b^1 = W_b^{-1}(\ell^J - \ell^{\bar{\psi}} M_{\psi\psi}^{-1} M_{\psi J})$$

and

$$\delta b^2 = W_b^{-1} \ell^{\bar{\psi}} M_{\psi\psi}^{-1} (\Delta\bar{\psi} + M_{\psi\phi})$$

and form

$$u^{(J+1)}(x) = u^{(J)}(x) - \frac{1}{2\gamma_0} \delta u^1(x) + \delta u^2(x)$$

and

$$b^{(J+1)} = b^{(J)} - \frac{1}{2\gamma_0} \delta b^1 + \delta b^2$$

Step 9. If all constraints are satisfied and $\delta u^1(x)$ and δb^1 are sufficiently small, terminate. Otherwise, return to Step 2.

The results of Theorem 8-2 hold for the optimal perturbations, and it may be shown that a necessary condition for convergence to a local optimum is $\delta u^1(x)$ and δb^1 approach zero. Discussions of Chapter 8 on use of the algorithm apply. They will not be repeated here.

9-3 A MINIMUM WEIGHT COLUMN

A minimum weight column problem has been solved in pars. 5-4 and 7-2 to illustrate the use of two optimization techniques. The same problem is solved in this paragraph, by the method of steepest descent, to illustrate the direct application of this technique to optimal structural design. The mathematical formulation of the problem is given in par. 7-2 and will be used here with a change in notation to be consistent with par. 9-2.

Fixing cross-sectional geometry and allowing cross-sectional area $u(x)$ to vary, as in par. 7-2, yields

$$I(x) = \alpha u^2(x) \tag{9-35}$$

The optimization problem is to choose $u(x)$, $0 < x < L$ to minimize

$$J = \int_0^L u(x)dx \tag{9-36}$$

subject to the constraints

$$\psi \equiv P_0 - P < 0 \tag{9-37}$$

$$\phi \equiv P/u(x) - \sigma_{max} < 0 \tag{9-38}$$

and

$$K(u)y \equiv \frac{d^2 y}{dx^2} = -P \frac{1}{E\alpha u^2} y \equiv PM(u)y \tag{9-39}$$

$$Cz \equiv \begin{bmatrix} y(0) \\ \frac{dy}{dx}(L) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \tag{9-40}$$

where the coordinate system is as shown in Fig. 7-1 with y replacing x and x replacing t.

In its present form, the boundary-value problem is just as in Eqs. 9-5 and 9-6 and is self-adjoint, so $y = y$ in par. 9-2. For use in the steepest-descent algorithm, Eq. 9-23 is

$$\Lambda^J = 1 \tag{9-41}$$

and Eq. 9-25 is

$$\Lambda^{\psi} = -\frac{3y^2(x)}{E\alpha u^3(x)} \Big/ \int_0^L \left[\frac{y^2(x)}{E\alpha u^2(x)}\right] dx \tag{9-42}$$

The computational algorithm of par. 9-2 now applies directly to the present problem. The solution of the eigenvalue problem, Eqs. 9-39 and 9-40, was obtained using the finite element analysis technique outlined in par. 5-4. The solution of this continuous problem required approximately the same time per iteration as the discrete technique but fewer iterations were generally required for convergence. Exactly the same results as given in par. 7-2 were obtained. The reader is referred to that paragraph for a tabulation of results.

9-4 A MINIMUM WEIGHT VIBRATING BEAM

As was pointed out in par. 7-2, the paper by Keller (Ref 11, Chapter 7) in 1960 presented a mathematically elegant method of designing the minimum weight column. The same method was applied by Niordson (Ref. 12, Chapter 7) in 1965 to find the simply supported beam of maximum natural frequency for a given volume of material in the beam. This method of solution resulted in a horribly nonlinear differential equation with serious singularities. While a solution was obtained for the vibrating beam problem, it is doubtful that the method could be extended for the solution of multimember structural design problems. The methods of Chapter 8, on the other hand, are quite general and will be used in this paragraph to routinely solve a minimum weight beam design problem with constraints on natural frequency.

Specifically, the problem considered here is the determination of the distribution of material along the centerline of a simply supported beam (see Fig. 9-2) so that the beam will be as light as possible and still have its fundamental frequency at least as large as a predetermined frequency ω_0. Further, so that the beam can support a minimum level of bending moment, it is required that the second moment of its cross-sectional area shall always be at least as large as a positive constant I_0.

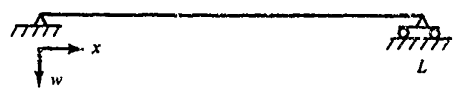


Figure 9-2. Simply Supported Vibrating Beam

As in the column problem of the preceding paragraph and par. 7-2, the geometry of the cross section is fixed and all dimensions are allowed to vary proportionally. If the area is denoted $u(t)$, then

$$I(t) = \alpha u^2(t) \tag{9-43}$$

where α is the minimum second moment of a cross section with the given geometry and unit area.

Since the material is to be specified with constant density, minimum weight is equivalent to minimum volume. The quantity to be minimized is, therefore,

$$J = \int_0^L u(t)dt\,. \tag{9-44}$$

The constraint on $I(t)$ discussed previously can now be written as

$$\phi = I_0 - \alpha u^2(t) \leq 0 \tag{9-45}$$

where $I_0 > 0$ is given.

The remaining feature of the problem to be accounted for is the constraint on natural frequency If the beam with $\alpha u^2(t) = I_0$ has a fundamental natural frequency of ω_0 or higher, then this is clearly the optimum beam. On the other hand, if this beam has a natural frequency below ω_0, then there must be points along the beam for which $\alpha u^2(t) > I_0$ and a meaningful design problem exists. The inequality on natural frequency is

$$\omega_0 - \omega < 0 \tag{9-46}$$

There are several ways in which the natural frequency of vibration of a beam may be related to the design of the beam $u(t)$. The relationship chosen here is the boundary-value problem describing lateral displacement during oscillation. It is given in Ref. 3 as

$$\left.\begin{aligned} &\frac{d^2}{dx^2}\left(E\alpha u^2 \frac{d^2 w}{dx^2}\right) = \rho\omega^2 uw \\ &w(0) = w(L) = 0 \\ &w''(0) = w''(L) = 0 \end{aligned}\right\} \tag{9-47}$$

where prime denotes differentiation with respect to x.

In order to put the boundary-value problem, Eq. 9-47, into the form Eqs. 9-1 and 9-2, define $y_1 = w$, $y_2 = E\alpha u^2(d^2y_1/dx^2)$, and $\rho\omega^2 = \zeta$. The problem, Eq. 9-47, is then

$$Ly \equiv \begin{bmatrix} \dfrac{d^2 y_2}{dx^2} \\ \\ \dfrac{d^2 y_1}{dx^2} - \dfrac{y^2}{E\alpha u^2} \end{bmatrix} = \zeta \begin{bmatrix} 0 \\ \\ uy_1 \end{bmatrix} \equiv \zeta My \tag{9-48}$$

with boundary conditions

$$\left.\begin{aligned} y_1(0) = y_1(L) = 0 \\ y_2(0) = y_2(L) = 0 \end{aligned}\right\} \tag{9-49}$$

The boundary-value problem, Eqs. 9-48 and 9-49, is self-adjoint so $y = \bar{y}$ in par. 9-2. The optimal design problem is well-defined and the notation of par. 9-2 applies directly. From Eq. 9-23,

$$\Lambda^J = 1$$

and from Eq. 9-25,

$$\Lambda^\psi = -\left(\frac{2y_2^2}{E\alpha u^3} - \zeta y_1^2\right) \bigg/ \int_0^L uy_1^2\, dx\,.$$

The computational steepest-descent algorithm may now be implemented in a direct manner.

As a numerical example, the given problem was solved with the data $E = 3 \times 10^7$ psi, $L = 10$ in., $\alpha = 1.0$, and $\rho = 0.00208$ slug/in^3 The eigenvalue problem was solved through use of a finite element structural analysis program. Even though there was no attempt at making the computational routines efficient, only 7 sec per iteration on an IBM 360-65 Computer were required. For most natural frequencies, 10 to 15 iterations were sufficient for convergence to within numerical accuracy of the computations. Results for a range of natural frequencies are given in Table 9-1. The general shapes of profiles of several of the optimum beams are shown in Fig. 9-3 to illustrate the optimum distribution of material.

9-5 A MINIMUM WEIGHT VIBRATING FRAME

The distribution of material along members of the frame shown in Fig. 9-4 is to be

TABLE 9-1

COMPARISON OF OPTIMAL BEAMS

Frequency, rad/sec	Volume of Uniform Beam of Length 10 in.*	Optimal Volume, in.^{3}	Weight Reduction or Material Savings, %
3600	0.9353	0.8967	4.13
1000	1.1546	1.0583	8.34
4400	1.3971	1.2536	10.27
4800	1.6627	1.4740	11.35
5200	1.9514	1.7189	11.92
5600	2.2631	1.9847	12.30
6000	2.5980	2.2705	12.61
10000	7.2165	6.3172	12.46

*Uniform beam of lowest volume having requi. d natural frequency

determined so that the frame is as lightweight as possible and has a fundamental natural

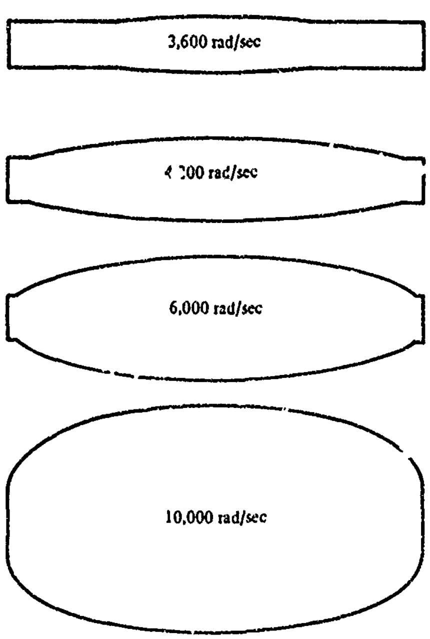


Figure 9-3. Profile of Optimal Beam

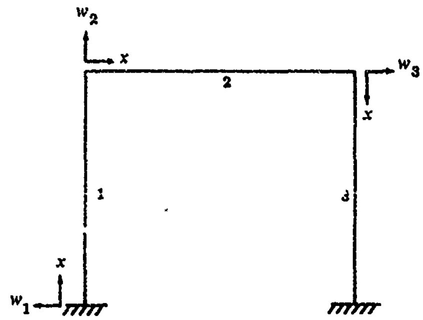


Figure 9-4. Portal Frame

frequency greater than or equal to a given ω_0. Further, as a form of strength requirement $I(x) > I_0 > 0$ is required.

For convenience, all members have the same length and all cross sections have the same geometry but may be scaled by a factor that varies with x. In this case, the area of cross sections $u_i(x)$, $i = 1,2,3$, uniquely determine the design of the beams when the beam material is chosen. Further, the second moment of the cross-sectional areas are

$$I_i(x) = \alpha_i \, u_i^2(x)$$

where α_i is a constant depending on the cross-sectional geometry chosen.

Defining

$$y_1 = w_1$$

$$y_2 = EI_1 w_1''$$

$$y_3 = w_2$$

$$y_4 = EI_2 w_2''$$

$$y_5 = w_3$$

$$y_6 = EI_3 w_3''$$

the differential equations for vibration of the frame are

$$Ky \equiv \begin{bmatrix} y_2'' \\ y_1'' - \dfrac{1}{EI_1} y_2 \\ y_4'' \\ y_3'' - \dfrac{1}{EI_2} y_4 \\ y_6'' \\ y_5'' - \dfrac{1}{EI_3} y_6 \end{bmatrix} = \rho\omega^2 \begin{bmatrix} u_1 y_1 \\ 0 \\ u_2 y_3 \\ 0 \\ u_3 y_5 \\ 0 \end{bmatrix} \equiv \zeta My \tag{9-50}$$

where $\zeta = \rho\omega^2$. Boundary conditions are

$$\left.\begin{aligned} &y_1(0) = 0 && y_1(L) = -y_5(0) \\ &y_1'(0) = 0 && y_1'(L) = y_3'(0) \\ &y_3(0) = 0 && y_3'(L) = y_5'(0) \\ &y_3(L) = 0 && y_2(L) = y_4(0) \\ &y_5(L) = 0 && y_4(L) = y_6(L) \\ &y_5'(L) = 0 \\ &y_2'(L) + y_6'(0) = \\ &\qquad -\left[\zeta \int_0^L u_2(x)dx\right] y_1(L) \end{aligned}\right\} \tag{9-51}$$

The boundary-value problem, Eqs. 9-50 and 9-51, is written in self-adjoint form. The last boundary condition in Eq. 9-51 is just Newton's second law applied to horizontal motion of Member 2. This boundary condition does not fit Eq. 9-6 exactly due to dependence on the design variable $u_2(x)$. It will have to be treated as a special case according to the comment following Eq. 9-13.

The perturbation boundary conditions from Eq. 9-51 are

$$\left.\begin{aligned} &\delta y_1(0) = 0 && \delta y_1(L) = -\delta y_5(0) \\ &\delta y_1'(0) = 0 && \delta y_1'(L) = \delta y_3'(0) \\ &\delta y_3(0) = 0 && \delta y_3'(L) = \delta y_5'(0) \\ &\delta y_3(L) = 0 && \delta y_2(L) = \delta y_4(0) \\ &\delta y_5(L) = 0 && \delta y_4(L) = \delta y_6(L) \\ &\delta y_5'(L) = 0 \\ &\delta y_2'(L) + \delta y_6'(0) = \\ &\qquad -\left[\zeta \int_0^L \delta u_2(x)dx\right] y_1(L) \\ &\qquad -\left[\zeta \int_0^L u_2(x)dx\right] \delta y_1(L) \end{aligned}\right\} \tag{9-52}$$

Two integrations by parts and elimination of boundary terms through use of Eqs. 9-51 and 9-52 yield

$$\int_0^L \delta y^T Ky \; dx = \int_0^L y^T K\delta y \; dx - \left[\zeta \int_0^L \delta u_2(x)dx\right] y_1(L) \tag{9-53}$$

Since the boundary value problem, Eqs. 9-50 and 9-51, is self-adjoint, $\bar{y} = y$ in the general formulation. The derivation of Eq. 9-21 holds and Eq. 9-53 may be substituted along with

$$\int_0^L y^T M\delta y \; dx = \int_0^L \delta y^T My \; dx$$

to obtain

$$\left\{ \int_0^L y^T M y \, dx \right\} \delta\zeta = \int_0^L \delta y^T [Ky - \zeta My] dx + \left[\zeta \int_0^L \delta u_2(x) dx \right] y_1(L) + \int_0^L \left\{ \left[\frac{2y_2^2}{E\alpha_1 u_1^3} \right] \delta u_1 + \left[\frac{2y_4^2}{E\alpha_2 u_2^3} \right] \delta u_2 + \left[\frac{2y_6^2}{E\alpha_3 u_3^3} \right] \delta u_3 \right\} dx$$

Solving for $\delta\zeta$,

$$\delta\zeta = \left[\frac{1}{\int_0^L (u_1 y_1^2 + u_2 y_3^2 + u_3 y_5^2) dx} \right] \times \int_0^L \left\{ \left[\frac{2y_2^2}{E\alpha_1 u_1^3} \right] \delta u_1 + \left[\frac{2y_4^2}{E\alpha_2 u_2^3} + \zeta y_1(L) \right] \delta u_2 + \left[\frac{2y_6^2}{E\alpha_3 u_3^3} \right] \delta u_3 \right\} dx \qquad (9\text{-}54)$$

By making the obvious choice for Λ^ζ, Eq. 9-54 can be written

$$\delta\zeta = \int_0^L \Lambda^{\zeta T} \delta u \, dx \; . \qquad (9\text{-}55)$$

This is precisely the form of Eq. 9-27, and the remainder of the general derivation of the steepest descent algorithm is valid. It should be noted that this derivation is formal, and rigorous verification of Eq. 9-54 is expected to be extremely difficult.

The algorithm of par. 9-2 was used to solve this problem in a direct manner. The eigenvalue problem was solved approximately by a finite element method. Data for this problem are $\alpha = 0.07958$, $\rho = 2.616 \times 10^4$ lb-sec^2/in.^{4}, $E = 10.3 \times 10^6$ psi, $I_0 = 0.009825$ in.^{4}, and $L = 10.0$ in. Weights of optimum frames are given in Table 9-2 for several frequency requirements, and the profile of an optimum frame is shown in Fig. 9-5.

TABLE 9-2

WEIGHT OF OPTIMUM FRAMES

ω_0, rad/sec	2000	3000	4000	5000
Optimum Weight, lb	1.73	2.56	3.59	4.69

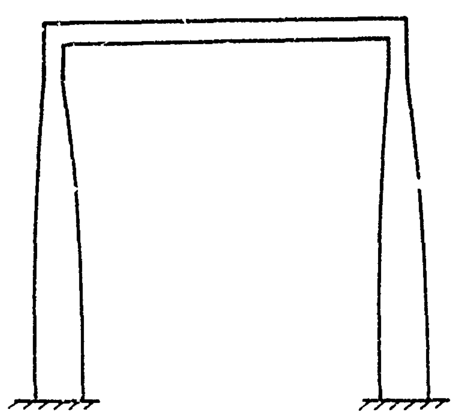


Figure 9-5. Profile of Optimum Frame

9-8 A MINIMUM WEIGHT FRAME WITH MULTIPLE FAILURE CRITERIA

In most structural design problems, constraints of several kinds must be treated simultaneously. In the present problem, constraints on member size, deflection, and buckling are enforced in the minimum weight design of the frame shown in Fig. 9-6. Area is

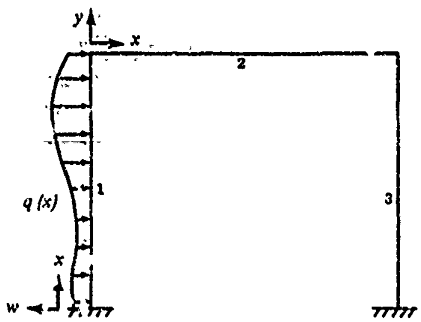


Figure 9-6. Laterally Loaded Frame

allowed to vary along the length of the first and second members but the geometrical shape of the cross section is fixed. Thus,

$$I_i(x) = \alpha_i u_i^2(x)$$

where α_i depends on the cross-sectional geometry and $u_i(x)$ is the cross-sectional area of the ith member at the point x. The size of the third member is fixed and all members are taken the same length.

Free body diagrams of the members are shown in Fig. 9-7.

The third member is uniform with constant modulus EI_3. Further, axial deformation of the second member is neglected so the top of

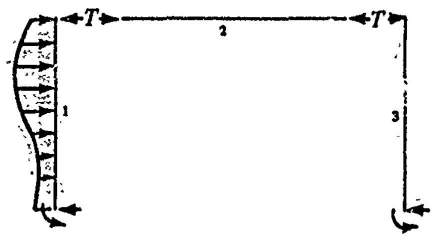


Figure 9-7. Free Bodies

the third member moves $-w(L)$ units to the right. Thus, by elementary beam theory (Ref. 3)

$$T = -\frac{3w(L)EI_3}{L^3} \tag{9-56}$$

The differential equation of deformation of the first member is

$$[E\alpha_1 u_1^2(x) w'']'' = -q(x) \tag{9-57}$$

and the boundary conditions are

$$\left.\begin{aligned} &w(0) = 0 \\ &w'(0) = 0 \\ &w''(L) = 0 \\ &-(E\alpha_1 u_1 w'')'(L) = \frac{3EI_3 w(L)}{L_3} \end{aligned}\right\} \tag{9-58}$$

To get the boundary-value problem, Eqs. 9-57 and 9-58. into the form of Eq. 9-1, define

$$\left.\begin{aligned} &z_1 = w \\ &z_2 = E\alpha_1 u_1 w'' = E\alpha_1 u_1^2 z_1'' \end{aligned}\right\} \tag{9-59}$$

The boundary-value problem, Eqs. 9-57 and 9-58, is then

$$Lz \equiv \begin{bmatrix} z_2'' \\ z_1'' - [1/E\alpha_1 u_1^2]\, z_2 \end{bmatrix} = \begin{bmatrix} -q(x) \\ 0 \end{bmatrix} \equiv Q(x) \qquad (9\text{-}60)$$

and

$$Bz \equiv \begin{bmatrix} z_1(0) \\ z_1'(0) \\ z_2(L) \\ -z_2'(L) - \dfrac{3EI_3 z_1(L)}{L^3} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \equiv q \qquad (9\text{-}61)$$

The equations which determine buckling load P of the second member are

$$Ky \equiv y'' = P\left(\frac{1}{E\alpha_2 u_2^2}\right) y \equiv PMy \qquad (9\text{-}62)$$

and

$$Cy \equiv \begin{bmatrix} y(0) \\ y(L) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \qquad (9\text{-}63)$$

The objective in the design problem is to choose $u_1(x)$ and $u_2(x)$ to minimize the weight of the first two members,

$$J = \gamma \int_0^L [u_1(x) + u_2(x)]\, dx \qquad (9\text{-}64)$$

The constraints to be enforced are

$$\psi_1 = -\left(\frac{3EI_3}{L^3}\right) z_1(L) - P < 0 \qquad (9\text{-}65)$$

$$\psi_2 = -z_1(L) - S < 0 \qquad (9\text{-}66)$$

and

$$\phi_i = -\alpha_i u_i(x) + I_0 < 0 \quad i = 1,2 \qquad (9\text{-}67)$$

The constraint, Eq. 9-65, requires that the axial load T in the second member be less than or equal the buckling load. A limit S is placed on horizontal deflection of the top of the frame in Eq. 9-66. The constraints, Eq. 9-67, are included to insure that member cross section does not go to zero anywhere. A more realistic constraint would be on bending stress but this would require a constraint of the form of Eq. 8-6. This constraint will be included in subsequent work but will not be treated here.

The constraints, Eqs. 9-65 and 9-66, do not fit directly into the basic formulation of this text and require special treatment. The linearized forms of these constraints are

$$-\left(\frac{3EI_3}{L^3}\right) \delta z_1(L) - \delta P < \Delta\psi \qquad (9\text{-}68)$$

and

$$-\delta z_1(L) < \Delta\psi_2 \qquad (9\text{-}69)$$

It remains to obtain expressions for $\delta z_1(L)$ and δP explicitly in terms of $\delta u_i(x)$. From Eq. 9-21, δP may be expressed in terms of $\delta u_2(x)$.

In order to obtain an expression for $\delta z_1(L)$, a Green's identity similar to Eq. 9-14 is needed. Integration twice by parts of $\int_0^L \lambda^T L\delta z\, dx$ yields

$$\int_0^L \lambda^T L\delta z\, dx = \int_0^L \delta z^T L\lambda\, dx$$

$$= (\lambda_1 \delta z_2' - \lambda_1' \delta z_2 + \lambda_2 \delta z_1' - \lambda_2' \delta z_1)\Big|_0^L .$$

Choosing λ such that $L\lambda = 0$, substituting for $L\delta z$ from Eq. 9-13, and using the linearized boundary conditions of Eq. 9-61, this becomes

$$\int_0^L \left(\frac{-2z_2}{E\alpha_1 u_1^3}\right) \delta u_1 dx = - \lambda_1(L)\left(\frac{3EI_3}{L^3}\right) \delta z_1(L) - \lambda_1(0)\delta z_2'(0) + \lambda_1'(0)\delta z_2(0) + \lambda_2(L)\delta z_1(L) - \lambda_2(L)\delta z_1(L) \tag{9-70}$$

The adjoint variable is chosen to satisfy $L\lambda = 0$ but no boundary conditions have yet been specified. Choosing

$$\left.\begin{aligned} -\frac{3EI_3}{L^3}\lambda_1(L) - \lambda_2(L) &= 1 \\ \lambda_1(0) = \lambda_1'(0) = \lambda_2(L) &= 0 \end{aligned}\right\} \tag{9-71}$$

the identity, Eq. 9-70, becomes

$$\delta z_1(L) = \int_0^L \left[\frac{-2z_2(x)\lambda_2(x)}{E\alpha_1 u_1^3(x)}\right] \delta u_1(x)dx \tag{9-72}$$

Thus an explicit relationship between δu and $\delta z_1(L)$ has been found and may be substituted directly into the linear constraints, Eqs. 9-68 and 9-69. With proper choice of notation, these inequalities fit into the form of Eq. 9-33.

In this problem the differential equations, Eqs. 9-1 and 9-5, are formally self-adjoint so $L^* = L$, $K^* = K$, and $M^* = M$ in the general theory. Since the boundary-value problem, Eqs. 9-62 and 9-63, is self-adjoint, $\bar{y} = y$ and the computation required in Step 2 of the steepest descent algorithm is considerably reduced.

This problem is now easily put in the form of the problem of par. 9-2. It was solved by direct application of the algorithm of that paragraph with the data $S = 4$ in., $L = 100$ in., $E = 3.0\text{x}10^7$ psi, $\alpha = 0.07958$, $I_0 = 0.0147$ in.^{4}, and area of member 3 is 4.0 in.^{2} The volume of the optimum frame for several values of q is given in Table 9-3. The profile of an optimum frame is shown in Fig. 9-8.

TABLE 9-3

VOLUME OF OPTIMUM FRAME

q, lb/in.	10	15	20	25
Optimum Volume, in.^{3}	502.1	531.1	556.0	653.9

9-7 A MINIMUM WEIGHT VIBRATING PLATE

In order to illustrate the use of the steepest descent method in higher dimensional problems, a minimum weight vibrating plate problem will be solved. A rectangular plate, Fig. 9-9, is specified by its thickness function $h(x_1, x_2)$ over the plate. The object here is to choose $h(x_1, x_2)$ such that the weight of the plate is as small as possible subject to the constraint that the natural frequency of lateral vibration is greater than or equal to a given frequency ω. Further, due to applied loads, a constraint of the form is

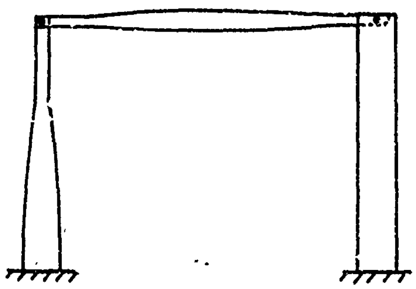


Figure 9-8. Profile of Minimum Weight Frame

$$h_0(x_1, x_2) - h(x_1, x_2) < 0 \qquad (9\text{-}73)$$

enforced. In the present problem, $h_0(x_1, x_2)$ is taken as a constant h_0.

Denoting bending moments (Ref. 4) by $y_1 = M_x$, $y_2 = M_y$, and $y_3 = M_{xy}$ and the lateral displacement by $y_4 = w$, the equations governing lateral vibration may be written (Ref. 4) in self-adjoint form as

$$K(h)y \equiv \begin{bmatrix} \dfrac{\partial^2 y_4}{\partial x_1^2} - \dfrac{12}{Eh^3}(y_1 - \mu y_2) \\[2ex] \dfrac{\partial^2 y_4}{\partial x_2^2} - \dfrac{12}{Eh^3}(y_2 - \mu y_1) \\[2ex] \dfrac{\partial^2 y_4}{\partial x_1 \partial x_2} - \left[\dfrac{6(1 + \mu)}{Eh^3}\right] y_3 \\[2ex] \dfrac{\partial^2 y_1}{\partial x_1^2} + \dfrac{\partial^2 y_2}{\partial x_2^2} + \dfrac{\partial^2 y_3}{\partial x_1 \partial x_2} \end{bmatrix} = \zeta \begin{bmatrix} 0 \\ 0 \\ 0 \\ hy_4 \end{bmatrix} \equiv \zeta M(h)y. \qquad (9\text{-}74)$$

where $\zeta = \rho\omega^2$ and ω is the natural frequency of vibration of the plate. The boundary-value problem for the simply supported plate with differential equations, Eqs. 9-74, is self-adjoint with the boundary conditions

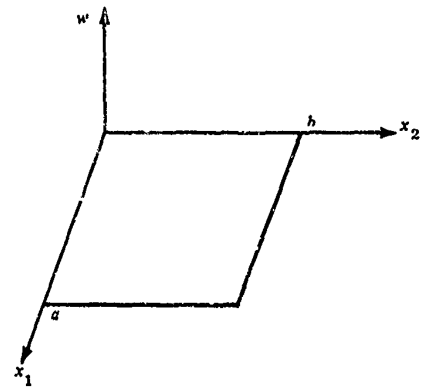


Figure 9-9. Simply Supported Plate

$$\left.\begin{aligned} &y_1(a, x_2) = 0, \; y_1(0, x_2) = 0 \\ &y_4(a, x_2) = 0, \; y_4(0, x_2) = 0 \\ &y_2(x_1, 0) = 0, \; y_2(x_1, b) = 0 \\ &y_4(x_1, 0) = 0, \; y_4(x_1, b) = 0 \end{aligned}\right\} \qquad (9\text{-}75)$$

The coordinate system and simply supported boundary of the plate are shown in Fig. 9-9.

To complete the formulation of this problem in terms of the preceding theory, a cost function is defined by

$$J = \int_0^b \int_0^a \gamma h(x_1, x_2) dx_1 dx_2 \qquad (9\text{-}76)$$

where γ is weight density of the plate material. The strength constraint in this problem is taken as Eq. 9-73 and the eigenvalue constraint is

$$\rho\omega_0^2 - \zeta < 0 \qquad (9\text{-}77)$$

This problem is now in the form of the general problem of par. 9-2. The domain Ω in this case is simply the rectangular region of the plate that is of dimension two. Further, since the boundary-value problem, Eqs. 9-74 and 9-75, is self-adjoint, $y = \bar{y}$ in the theory of par. 9-2.

Using the definition of K and M, $\phi \equiv h_0 - h(x_1,x_2)$ in Eq. 9-73, and $e_j(\zeta) \equiv \rho\omega_0^2 - \zeta$ in Eq. 9-77, from Eq. 9-23,

$$\Lambda^{\zeta} = \gamma \tag{9-78}$$

and from Eq. 9-25

$$\Lambda^{\psi l} = -\left[\frac{36(y_1^2 - 2\mu y_1 y_2 + y_2^2) + 18(1+\mu)y_3^2}{Eh^4} - \zeta y_4^2\right] \Bigg/ \int_0^b \int_0^a h y_4^2 \, dx_1 dx_2 \tag{9-79}$$

This problem is now in the form of the general problem of par. 9-2. It was solved by direct application of the algorithm of that paragraph. The eigenvalue and eigenfunction for the variable thickness plate were determined approximately by the Ritz technique (Ref. 5). Data for this problem are $a = b = 5.0$ in., $E = 3.0\text{x}10^7$ psi, $\rho = 7.45\text{x}10^{-4}$ lb-sec^2/in.^{4}, $\nu = 0.30$, and $\omega_0 = 1375$ rad/sec. A uniform plate with $\zeta = \rho\omega^2 = 1400$ was taken as the initial estimate. The volume of the optimum plate is 10.71 in.^{3} Double symmetry of the optimum plate was observed about the axes through $(a/2,b/2)$. One quarter of the optimum plate with contour lines is shown in Fig. 9-10.

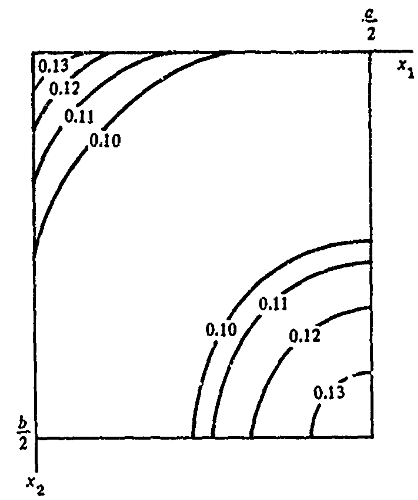


Figure 9-10. Contours of Optimum Plate

REFERENCES

1. C. Lanczos, *Linear Differential Operators*, Van Nostrand, London, 1961.
2. T. Kato, *Perturbation Theory for Linear Operators*, Springer-Verlay, New York, New York, 1966.
3. A. Higdon, E. H. Ohlsen, and W. B. Stiles, *Mechanics of Materials*.
4. C. T. Wang, *Applied Elasticity*, McGraw-Hill, New York, 1953.
5. S. G. Mikhlin, *Variational Methods in Mathematical Physics*, MacMillan, New York, 1964

APPENDIX A

CONVEXITY

Convex functions and sets as defined in Chapter 2 play an important role in optimization theory. It is generally possible to obtain much more comprehensive results in nonlinear programming problems that are convex than in the nonconvex case. Some of the more important results due to convexity are given in Chapter 4, par. 4-2.

In order that this appendix is self-contained, definitions of general convex sets and functions will be repeated here. For a complete treatment of convexity, the reader is referred to Ref. 1.

Definition A-1: Let D be a subset of R^n. D is called a convex set if for any points x and y in D, $x + \theta(y - x)$ is also in D for all θ such that $0 < \theta < 1$.

The collection of points $x + \theta(y - x)$, $0 < \theta < 1$, is just a straight line from x to y. Def. A-1 just says, then, that a set in R^n is convex if the straight line joining any pair of points in the set lies entirely in the set. For example, in R^2 (the plane) the set of points inside the unit circle is convex (see Fig. A-1(A)) whereas the star-shaped region in Fig. A-1(B) is not convex.

Convex functions have as their prototype $f(x) = x^2$ in R^1. The graph of this function is shown in Fig. A-2. Note in this figure that if a straight line is constructed between any two points $[z, f(z)]$ and $[y, f(y)]$, then this line is above the graph of $f(x)$ at all points between z

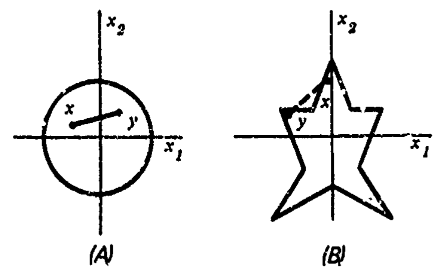


Figure A-1. Examples; Convex Case and Nonconvex Case

and y. This is precisely the property which characterizes convex functions. Analytically, this property is expressed by the inequality

$$f[z + \theta(y - z)] < f(z) + \theta[f(y) - f(z)]$$

for all θ with $0 < \theta < 1$.

The same idea holds in R^n where convex functions are characterized by

Definition A-2: Let the real valued function $f(x)$ be defined on the convex subset D

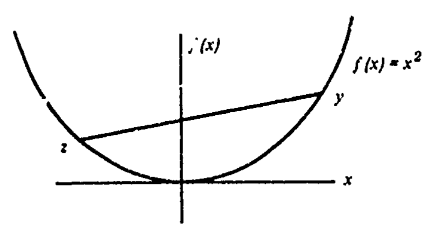


Figure A-2. Graph of $f(x) = x^2$ in R^1

of R^n. Then $f(x)$ is called a convex function if for any points z and y in D,

$$f[z + \theta(y - z)] < f(z) + \theta[f(y) - f(z)] \tag{A-1}$$

for all θ with $0 < \theta < 1$.

Convex functions and convex sets are related as is shown in

Theorem A-1: The set of points D in R^n which satisfy $g_i(x) < 0$, $i = 1, ..., m$, is convex if each of the functions $g_i(x)$ is convex in R^n.

One further property of convex functions is extremely important for applications. It is established by

Theorem A-2: If $f(x)$ is differentiable and convex on the convex subset D of R^n, then

$$f(x) > f(y) + \nabla f(y) \cdot (z - y) \tag{A-2}$$

for all z and y in D.

All these desirable properties of convex functions will go to waste unless one is able to test a given function for convexity. The following three theorems provide useful tests:

1. *Theorem A-3:* If $f(x)$ is twice continuously differentiable in a convex subset D of R^n, it is convex in D if and only if the quadratic form

$$S^T \left[\frac{\partial^2 f}{\partial x_i \partial x_j}\right] S \quad (\text{or } S^T \cdot \nabla^2 f \cdot S) \tag{A-3}$$

is positive-semidefinite at each point in D.

2. *Theorem A-4:* If the functions $q_i(x)$, $i = 1, ..., r$, are convex in the convex subset D of R^n and $\alpha_i > 0$, $i = 1, ..., r$, then

$$\sum_{i=1}^{r} \alpha_i q_i(x)$$

is convex in D.

3. *Theorem A-5:* If $g(x)$ is twice continuously differentiable, $g(x) < 0$, and $g(x)$ is convex; then, $1/[g(x)]$ is convex.

REFERENCE

1. R. T. Rockafeller, *Convex Analysis*, Princeton University Press, Princeton, N.J., 1970.

APPENDIX B

ANALYSIS OF BEAM-TYPE STRUCTURES

Finite and discrete element methods of structural analysis (Refs. 1,3), require a knowledge of the behavior of each element in the structure. Once each element is described, then the governing equations of the entire structure may be derived. Energy methods are generally used to obtain the governing equations.

B-1 ELEMENT ANALYSIS

In order to apply energy theorems for the analysis of a structure, the potential energy due to strain, kinetic energy, and change in external dimensions due to bending must be described. The basic idea is to assign generalized displacement functions, which are of the form expected in structural deformation and that are uniquely specified when the displacement of both ends of the beam is known. A typical beam with its deformation sign convention is shown in Fig. B-1. The displacement u_1, u_2, u_3, and u_4 are components of endpoint displacement and u_5 and u_6 are endpoint rotations.

The longitudinal displacement of a point x, $0 < x < \ell$, on the beam due to longitudinal strain is approximated by

$$s(x) = -u_1\left(\frac{x-\ell}{\ell}\right) + u_2\frac{x}{\ell}\,. \qquad \text{(B-1)}$$

Lateral displacement of the beam at a point x is approximated by

Figure B-1. Basic Beam Element

$$w(x) = \frac{u_3}{\ell^3}(2x^3 - 3\ell x^2 + \ell^3) - \frac{u_4}{\ell^3}(2x^3 - 3\ell x^2) + \frac{u_5}{\ell^2}(x^3 - 2\ell x^2 + \ell^2 x) + \frac{u_6}{\ell^2}(x^3 - \ell x_2). \qquad \text{(B-2)}$$

It should be stressed that the longitudinal displacement $s(x)$ is due only to longitudinal strain in the beam and not due to the change in length caused by the lateral displacement $w(x)$.

The potential energy PE due to deformation of the beam is (Ref. 2):

$$\begin{aligned} PE &= \frac{1}{2}\int_0^\ell AE\left(\frac{ds}{dx}\right)^2 dx + \frac{1}{2}\int_0^\ell EI\left(\frac{d^2w}{dx^2}\right)^2 dx \\ &= \frac{1}{2}\int_0^\ell AE\left(-\frac{u_1}{\ell}+\frac{u_2}{\ell}\right)^2 dx + \frac{1}{2}\int_0^\ell EI\left[\frac{u_3}{\ell^3}(12x-6\ell) - \frac{u_4}{\ell^3}(12x-6\ell) + \frac{u_5}{\ell^2}(6x-4\ell) + \frac{u_6}{\ell^2}(6x-2\ell)\right]^2 dx \end{aligned} \qquad \text{(B-3)}$$

which is

$$PE = \frac{1}{2} u^T \frac{E}{\ell^3} \begin{bmatrix} A\ell^2 & -A\ell^2 & 0 & 0 & 0 & 0 \\ -A\ell^2 & A\ell^2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 12I & -12I & -6I\ell & -6I\ell \\ 0 & 0 & -12I & 12I & 6I\ell & 6I\ell \\ 0 & 0 & -6I\ell & 6I\ell & 4I\ell^2 & 2I\ell^2 \\ 0 & 0 & -6I\ell & 6I\ell & 2I\ell^2 & 4I\ell^2 \end{bmatrix} u \tag{B-4}$$

where $u = [u_1, u_2, u_3, u_4, u_5, u_6]^T$.

Similarly, the kinetic energy KE of the beam is

$$KE = \frac{1}{2} \int_0^\ell \rho A \left(\frac{ds^2}{dt} + \frac{dw^2}{dt} \right) dx$$

$$= \frac{1}{2} \int_0^\ell \rho A \left\{ \left[-\dot{u}_1 \left(\frac{x-\ell}{\ell} \right) + \dot{u}_2 \frac{x}{\ell} \right]^2 \right.$$

$$+ \left[\frac{\dot{u}_3}{\ell^3} (2x^3 - 3\ell x^2 + \ell^3) \right.$$

$$- \frac{\dot{u}_4}{\ell^3} (2x^3 - 3\ell x^2) \tag{B-5}$$

$$+ \frac{\dot{u}_5}{\ell^2} (x^3 - 2\ell x^2 + \ell x)$$

$$\left. \left. + \frac{\dot{u}_6}{\ell^2} (x^3 - \ell x^2) \right]^2 \right\} dx$$

or

$$KE = \frac{1}{2} \dot{u}^T \frac{\rho A \ell}{420} \begin{bmatrix} 140 & 70 & 0 & 0 & 0 & 0 \\ 70 & 140 & 0 & 0 & 0 & 0 \\ 0 & 0 & 156 & 54 & -22\ell & 13\ell \\ 0 & 0 & 54 & 156 & -13\ell & 22\ell \\ 0 & 0 & -22\ell & -13\ell & 4\ell^2 & -3\ell^2 \\ 0 & 0 & 13\ell & 22\ell & -3\ell^2 & 4\ell^2 \end{bmatrix} \dot{u}. \tag{B-6}$$

The shortening of the beam $\Delta\ell$ due to the lateral displacement is

$$\Delta\ell = \ell - \int_0^\ell \left[1 - \left(\frac{dw}{dx} \right)^2 \right]^2 dx$$

$$\approx \int_0^\ell \frac{1}{2} \left(\frac{dw}{dx} \right)^2 dx$$

$$= \frac{1}{2} \int_0^\ell \left[\frac{u_3}{\ell^3} (6x^2 - 6\ell x) \right. \tag{B-7}$$

$$- \frac{u_4}{\ell^3} (6x^2 - 6\ell x)$$

$$+ \frac{u_5}{\ell^2} (3x^2 - 4\ell x + \ell^2)$$

$$\left. + \frac{u_6}{\ell^2} (3x^2 - 2\ell x) \right]^2 dx$$

or

$$\Delta\ell = u^T \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{3}{5\ell} & -\frac{3}{5\ell} & \frac{1}{20} & \frac{1}{20} \\ 0 & 0 & -\frac{3}{5\ell} & \frac{3}{5\ell} & -\frac{1}{20} & -\frac{1}{20} \\ 0 & 0 & \frac{1}{20} & -\frac{1}{20} & \frac{\ell}{15} & -\frac{\ell}{60} \\ 0 & 0 & \frac{1}{20} & -\frac{1}{20} & -\frac{\ell}{60} & \frac{\ell}{15} \end{bmatrix} u \tag{B-8}$$

B-2 VARIATIONAL PRINCIPLES

In most structural analysis work it is more convenient to use variational principles to describe the state of the structure than it is to use Newton's laws directly. For use in analysis of the structures considered here, the basic variational principles are stated. These principles apply to systems that are unconstrained in the particular coordinate system chosen; i.e., once the coordinates, u_i, $i = 1, ..., n$, that describe the state of the system are chosen, there are no algebraic relationships between these variables. Further, it is required that all force-state relationships for external or internal forces are continuous, i.e., small changes in state yield only small changes in forces.

Let W' denote the work done by all forces on a structure due to an admissible displacement. Then, the system is called conservative if the work W' done in any displacement that returns to the original state is zero; i.e., the system is conservative if no energy is required to change the state of the structure and bring it back to the original state along any path. A structure would be nonconservative, for example, if sliding friction or viscous damping were present.

Starting from some reference state of the structure u_0, define the stored energy in the structure at state u to be $U(u)$. Note that since the system is conservative, $U(u)$ depends only on u and not on the path the state variable followed in getting from u_0 to u. Likewise, the negative of work done by the external forces acting on the structure due to the change in state variables will be denoted $\Omega(u)$. Again, $\Omega(u)$ depends only on the final state and not on the path from u_0 to u. With this notation, the total potential energy $V(u)$ is defined as

$$V(u) = U(u) + \Omega(u) \tag{B-9}$$

Equilibrium states of a structure can now be characterized in terms of the total potential energy. For unconstrained conservative systems, a necessary and sufficient condition that u be an equilibrium state is that

$$\delta V(u) = \sum_{i=1}^{n} \frac{\partial V(u)}{\partial u_i} \delta u_i = 0 \tag{B-10}$$

for all δu_i. This is equivalent to

$$\frac{\partial V(u)}{\partial u_i} = 0, \; i = 1, ..., n. \tag{B-11}$$

This result is proved in Ref. 2, page 23.

A second result is that a conservative, unconstrained system is in stable equilibrium at a state u if and only if the total potential energy is a relative minimum at u. This result is proved in Ref. 2, page 30. For the kind of structural systems considered here, it is further shown, Ref. 2, page 211, that at a buckling load P_{cr}, the second variation must be positive semi-definite, i.e.,

$$\delta^2 V(u) = \delta u^T \frac{\partial^2 V}{\partial u^2}(u)\delta u \geq 0 \tag{B-12}$$

and further that there is a $\delta\bar{u}$ such that

$$\delta^2 V(u) = \delta\bar{u}^T \frac{\partial^2 V}{\partial u^2}(u)\delta\bar{u} = 0. \tag{B-13}$$

The inequality, Eq. B-12, shows that zero is a relative minimum and Eq. B-13 shows that relative minimum is attained for $\delta u = \delta\bar{u}$. For $\delta^2 V(u)$ treated as a function of δu, then, it is necessary that

$$\frac{\partial}{\partial \delta u_i}[\delta^2 V(u)] = 0, \; i = 1, ..., n$$

at $\delta u = \delta \bar{u}$. Thus

$$\frac{\partial^2 V}{\partial u^2}(u)\, \delta \bar{u} = 0 \qquad \text{(B-14)}$$

The condition, Eq. B-14, determines buckling loads for the structural system. All this analysis requires, of course, that $V(u)$ is at least twice continuously differentiable.

These laws are for static behavior of the structural system. Dynamically, the structure is governed by Lagrange's equations of motion. First, define the kinetic energy T as the quadratic form

$$T = \frac{1}{2}\dot{u}^T M \dot{u} \qquad \text{(B-15)}$$

where

$$M = [m_{ij}]$$

and the m_{ij} are generalized masses. The generalized mass matrix M is defined by the transformation

$$z = f(Z)$$

where

$$Z = [Z_1, ..., Z_p]$$

and Z_i are components of physical displacement of the masses of the structure. Therefore,

$$T = \frac{1}{2}\sum_{i=1}^{p} \dot{Z}_i^2\, m_i = \frac{1}{2}\sum_{i=1}^{p}\left[\left(\frac{\partial f}{\partial Z}\right)^{-1} \dot{u}\right]^T \times m_i \left[\left(\frac{\partial f}{\partial Z}\right)^{-1} \dot{u}\right]$$

$$= \frac{1}{2}\sum_{i=1}^{p} \dot{u}^T \left[\left(\frac{\partial f^T}{\partial Z}\right)^{-1} m_i \left(\frac{\partial f}{\partial Z}\right)^{-1}\right]\dot{u}$$

where m_i are element masses. In the notation of Eq. B-15,

$$M = \sum_{i=1}^{p}\left[\left(\frac{\partial f^T}{\partial Z}\right)^{-1} m_i \left(\frac{\partial f}{\partial Z}\right)^{-1}\right].$$

Putting

$$L = T - V \qquad \text{(B-16)}$$

Lagrange's equations of motion are simply (Ref. 2, page 239)

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{u}_i}\right) - \frac{\partial L}{\partial u_i} = 0, \; i = 1, ..., n. \qquad \text{(B-17)}$$

B-3 EQUATIONS OF STRUCTURAL ANALYSIS

The given variational principles may be applied to obtain the governing equations of structural analysis. The element properties described by Eqs. B-4, B-6, and B-8 may be used to generate the potential and kinetic energy of the entire structure. From the definitions, Eqs. B-9 and B-15,

$$V = \sum_j PE^j - \sum_k (u^k + \bar{u}^k)\, F^k \qquad \text{(B-18)}$$

and

$$T = \sum_j KE^j \qquad \text{(B-19)}$$

where superscript j denotes the jth element of the structure, k denotes the components of displacement at joints, F^k is the component of external force corresponding to u^k, and $\bar{u}^k$ is the displacement in the same direction as u^k but due to the $\Delta \ell$ components of bar deformation. The displacements $\bar{u}^k$ will be determined from structure geometry and the $\Delta\ell^j$ of individual members given in Eq. B-8. For determining the equilibrium equations B-11, the displacements $\bar{u}^k$ are generally neglected since they are at least quadratic in u so they will be small if no buckling occurs. It is just these quadratic terms, however, that predict buckling behavior of structures.

If a composite displacement vector u is formed from the components of all the member displacements, then matrices K and M may be defined by

$$\frac{1}{2} u^T K u = \frac{1}{2} \sum_j u^{j^T} K^j u^j$$

and

$$\frac{1}{2} \dot{u}^T M \dot{u} = \frac{1}{2} \sum_j \dot{u}^{j^T} M^j \dot{u}^j$$

where summation is taken over all elements of the structure and K^j and M^j are defined in Eqs. B-4 and B-5. The equations for displacement, buckling, and dynamic motion may now be determined directly from Eqs. B-11, B-14, and B-17.

REFERENCES

1. O. C. Zienkiewicz, *The Finite Element Method*, McGraw-Hill, London, 1967.

2. H. L. Langhaar, *Energy Methods in Applied Mechanics*, Wiley, New York, 1962.

3. J. S. Przemieniecki, *Theory of Matrix Structural Analysis*, McGraw-Hill, New York, 1968.

INDEX

INDEX (Con't.)

INDEX (Con't.)

(AMCRD-TV)

FOR THE COMMANDER:

OFFICIAL:

JOHN LYCAS
Colonel, GS
Chief, HQ Admin Mgt Ofc

JOSEPH W. PEZDIRTZ
Major General, USA
Chief of Staff

DISTRIBUTION:
Special

★ U. S. GOVERNMENT PRINTING OFFICE 1973 O 540 802 (5010A

ENGINEERING DESIGN HANDBOOKS

Available to AMC activities, DOD agencies, and Government agencies from Letterkenny Army Depot, Chambersburg, PA 17201. Available to contractors and universities from National Technical Information Service (NTIS), Department of Commerce, Springfield, VA 22151 EXCEPT WHERE NOTED.

No. AMCP 706-	Title
100	Design Guidance for Producibility
104	Value Engineering
106#	Elements of Armament Engineering, Part One, Sources of Energy
107#	Elements of Armament Engineering, Part Two, Ballistics
108#	Elements of Armament Engineering, Part Three, Weapon Systems and Components
109	Tables of the Cumulative Binomial Probabilities
110	Experimental Statistics, Section 1, Basic Concepts and Analysis of Measurement Data
111	Experimental Statistics, Section 2, Analysis of Enumerative and Classificatory Data
112	Experimental Statistics, Section 3, Planning and Analysis of Comparative Experiments
113	Experimental Statistics, Section 4, Special Topics
114	Experimental Statistics, Section 5, Tables
115	Environmental Series, Part One, Basic Environmental Concepts
116	*Environmental Series, Part Two, Basic Environmental Factors
120	Criteria for Environmental Control of Mobile Systems
121	Packaging and Pack Engineering
123	Hydraulic Fluids
125	Electrical Wire and Cable
127	Infrared Military Systems, Part One
128(S)#	Infrared Military Systems, Part Two (U)
130	Design for Air Transport and Airdrop of Materiel
132	*Maintenance Engineering
133	*Maintainability Engineering Theory and Practice
134	Maintainability Guide for Design
135	Inventions, Patents, and Related Matters
136	Servomechanisms, Section 1, Theory
137	Servomechanisms, Section 2, Measurement and Signal Converters
138	Servomechanisms, Section 3, Amplification
139	Servomechanisms, Section 4, Power Elements and System Design
140	Trajectories, Differential Effects, and Data for Projectiles
150	Interior Ballistics of Guns
160(S)#	Elements of Terminal Ballistics, Part One, Kill Mechanisms and Vulnerability (U)
161(S)#	Elements of Terminal Ballistics, Part Two, Collection and Analysis of Data Concerning Targets (U)
162(SRD)#	Elements of Terminal Ballistics, Part Three, Application to Missile and Space Targets (U)
165	Liquid-Filled Projectile Design
170(C)#	**Armor and Its Applications (U)
175#	Solid Propellants, Part One
176(C)#	Solid Propellants, Part Two (U)
177	Properties of Explosives of Military Interest
178(C)	+Properties of Explosives of Military Interest, Section 2 (U) (REPLACED BY -177)
179#	**Explosive Trains
180	Principles of Explosive Behavior
181	*Explosions in Air, Part One
182(S)#	*Explosions in Air, Part Two (U)
185#	Military Pyrotechnics, Part One, Theory and Application
186	Military Pyrotechnics, Part Two, Safety, Procedures and Glossary
187#	Military Pyrotechnics Part Three, Properties of Materials Used in Pyrotechnic Compositions
188#	*Military Pyrotechnics, Part Four, Design of Ammunition for Pyrotechnic Effects
189	Military Pyrotechnics, Part Five, Bibliography
190	*Army Weapon System Analysis
191	System Analysis and Cost-Effectiveness
195	*Development Guide for Reliability, Part One, Introduction, Background, and Planning for Army Materiel Requirements
196	*Development Guide for Reliability, Part Two, Design for Reliability
197	*Development Guide for Reliability, Part Three, Reliability Prediction
198	*Development Guide for Reliability, Part Four, Reliability Measurement
199	*Development Guide for Reliability, Part Five, Contracting for Reliability
200	*Development Guide for Reliability, Part Six, Mathematical Appendix and Glossary

No. AMCP 706-	Title
201	*Helicopter Engineering, Part One, Preliminary Design
202	*Helicopter Engineering, Part Two, Detail Design
203	Helicopter Engineering, Part Three, Qualification Assurance
204	*Helicopter Performance Testing
205	*Timing Systems and Components
210	Fuzes
211(C)#	Fuzes, Proximity, Electrical, Part One (U)
212(S)#	Fuzes, Proximity, Electrical, Part Two (U)
213(S)#	Fuzes, Proximity, Electrical, Part Three (U)
214(S)#	Fuzes, Proximity, Electrical, Part Four (U)
215(C)#	Fuzes, Proximity, Electrical, Part Five (U)
235	Hardening Weapon Systems Against RF Energy
238	*Recoilless Rifle Weapon Systems
239	*Small Arms Weapon Systems
240(S)#	Grenades (U)
242	Design for Control of Projectile Flight Characteristics (REPLACES -246)
244	Ammunition, Section 1, Artillery Ammunition--General, with Table of Contents, Glossary, and Index for Series
245(C)#	Ammunition, Section 2, Design for Terminal Effects (U)
246	+Ammunition, Section 3, Design for Control of Flight Characteristics (REPLACED BY -242)
247#	Ammunition, Section 4, Design for Projection
248	+Ammunition, Section 5, Inspection Aspects of Artillery Ammunition Design
249	Ammunition, Section 6, Manufacture of Metallic Components of Artillery Ammunition
250	Guns--General
251	Muzzle Devices
252	Gun Tubes
253	*Breech Mechanism Design
255	Spectral Characteristics of Muzzle Flash
260	Automatic Weapons
270	**Propellant Actuated Devices
280	Design of Aerodynamically Stabilized Free Rockets
281(SRD)#	Weapon System Effectiveness (U)
282	+Propulsion and Propellants (REPLACED BY -285)
283	Aerodynamics
284(C)#	Trajectories (U)
285	Elements of Aircraft and Missile Propulsion (REPLACES -282)
286	Structures
290(C)#	Warheads--General (U)
291	Surface-to-Air Missiles, Part One, System Integration
292	Surface-to-Air Missiles, Part Two, Weapon Control
293	Surface-to-Air Missiles, Part Three, Computers
294(S)#	Surface-to-Air Missiles, Part Four, Missile Armament (U)
295(S)#	Surface-to-Air Missiles, Part Five, Countermeasures (U)
296	Surface-to-Air Missiles, Part Six, Structures and Power Sources
297(S)#	Surface-to-Air Missiles, Part Seven, Sample Problem (U)
327	Fire Control Systems--General
329	Fire Control Computing Systems
331	Compensating Elements
335(SRD)#	*Design Engineers' Nuclear Effects Manual, Volume I, Munitions and Weapon Systems (U)
336(SRD)#	*Design Engineers' Nuclear Effects Manual, Volume II, Electronic Systems and Logistical Systems (U)
337(SRD)#	*Design Engineers' Nuclear Effects Manual, Volume III, Nuclear Environment (U)
338(SRD)#	*Design Engineers' Nuclear Effects Manual, Volume IV, Nuclear Effects (U)
340	Carriages and Mounts--General
341	Cradles
342	Recoil Systems
343	Top Carriages
344	Bottom Carriages
345	Equilibrators
346	Elevating Mechanisms
347	Traversing Mechanisms
350	Wheeled Amphibians
355	The Automotive Assembly
356	Automotive Suspensions
357	Automotive Bodies and Hulls
360	*Military Vehicle Electrical Systems
445	Sabot Technology Engineering

*UNDER PREPARATION--not available
+OBSOLETE--out of stock
**REVISION UNDER PREPARATION
#NOT AVAILABLE FROM NTIS