(*AI EDAM*) (1987) **1**(3), 169–189

INNOVATIVE DESIGN OF MECHANICAL STRUCTURES FROM FIRST PRINCIPLES

JONATHAN CAGAN AND ALICE M. AGOGINO
Intelligent Systems Research Group, Department of Mechanical Engineering, University of California at Berkeley, Berkeley CA 94720, *U.S.A.*

In this paper a unique design methodology known as 1stPRINCE (FIRST PRINciple Computational Evaluator) is developed to perform innovative design of mechanical structures from first principle knowledge. The method is based on the assumption that *the creation of innovative designs of physical significance, concerning geometric and material properties, requires reasoning from first principles.* The innovative designs discovered by 1stPRINCE differ from routine designs in that new primitives are created. Monotonicity analysis and computer algebra are utilized to direct design variables in a globally optimal direction relative to the goals specified. In contrast to strict constraint propagation approaches, formal qualitative optimization techniques efficiently search the solution space in an optimizing direction, eliminate infeasible and suboptimal designs, and reason with both equality and inequality constraints. Modification of the design configuration space and the creation of new primitives, in order to meet the constraints or improve the design, are achieved by manipulating mathematical quantities such as the integral. The result is a design system which requires a knowledge base only of fundamental equations of deformation with physical constraints on variables, constitutive relations, and fundamental engineering assumptions; no pre-compiled knowledge of mechanical behavior is needed. Application of this theory to the design of a beam under torsion leads to designs of a hollow tube and a composite rod exhibiting globally optimal behavior. Further, these optimally-behaved designs are described symbolically as a function of the material properties and system parameters. This method is implemented in a LISP environment as a module in a larger intelligent CAD system that integrates qualitative, functional and numerical computation for engineering applications.

1. Introduction

Artificial intelligence techniques have recently entered the mechanical engineering discipline through numerous applications. Implementations in the areas of design and analysis, however, have been few. We classify three levels of design similar to Brown and Chandrasekaran (1986): routine, innovative, creative. In routine design, existing parameters are varied until a satisfactory design is found. Most design systems presently work within this design level. PRIDE (Mittal *et al.*, 1985) designs transport systems in copiers for the Xerox Corporation. Using object-oriented programming and constraint propagation techniques, the system selects and modifies parts until constraints are met. SACON (Bennett and Engelmore, 1979) and, more recently, PLASHTRAN (Cagan and Genberg, 1987) aid users through the application of finite element analysis codes to model physical structures; the former is based on the EMYCIN goal-directed, rule-based technology, and the latter is implemented in LOOPS (Xerox Corporation, 1983) for a frame-based, data-driven system. DOMINIC I (Dixon *et al.*, 1987) describes a design procedure in which expert systems performing V-belt design and heat fin design can be developed. Another architecture for routine design is DSPL (Design Specialists and Plans Language), as described in Brown and Chandrasekaran (1986).

The above programs do not reason from first principles. By 'first principles' we mean the fundamental physical equations and concepts utilized by engineers. PRIDE is not concerned with deformation of parts such as a pinch roller and does not require analysis from first principles. Systems such as SACON and PLASHTRAN are capable of solving problems within their specified domain. If, however, a structure is described outside of the system domain, the code may give incorrect responses. Systems derived from DOMINIC use a hill-climbing technique to vary parameters within a design space for redesign applications.

Reasoning in typical expert systems is accomplished by rules and inheritance. Their only connection to the physical problem is through terminology and variable bindings. In PLASHTRAN, if the user models a shell, the system binds the variable *type.of.structure* to the term 'shell'. It can then refer to and recognize

0890–0604/87/030169 + 21$02.00/0

(through pattern recognition) the word 'shell', but the computer does not actually reason about the physical properties and behaviors of structural shells. Further, the system cannot make creative design decisions or structural modifications.

Innovative design is the process of deriving new design features from previous designs. Other approaches to innovative design systems include Ulrich and Seering (1987) who utilize 'novel combination', the process of extracting individual attributes from known devices and combining them together to form new devices, to creative new fastener designs. The process of novel combination extracts features while propagating constraints, but doesn't consider the physical relationships between parts. PROMPT (Murthy and Addanki, 1987) innovates structural designs and will be discussed in detail below. EDISON (Dyer *et al.*, 1986) is a design system which utilizes naive physical relationships and qualitative reasoning with planning and heuristics on discovery/invention. In EDISON, devices are mutated from a limited library of standard devices, such as a can opener and a door, with various functional purposes of process to satisfy needs of a different process. Lenat (1983) presents EURISKO, a domain-independent discovery program that learns new heuristics. One application of EURISKO is the design of a fleet of ships which won the Trillion Credits Squadron national tournament. Again, in EDISON and EURISKO, physical relationships between parts are not considered.

In the third level of design, creative design, new primitives which have no obvious relationship to previous configurations are created. Creativity is a complex cognitive process that is not well understood in humans and thus is not easily codifiable in computer-based systems. Coyne *et al.* (1987) examine the creative potential of knowledge-based systems from a 'narrow information processing perspective'. They propose that criteria for evaluating the creativity of computer-based systems should include the ability to acquire knowledge, control internal processes and change internal structure. Only limited attempts at creative knowledge-based systems are described.

Qualitative process theory and naive physics are approaches to reason about the physical world from first principles as described by Hayes (1985), Forbus (1983), DeKleer and Brown (1983), and Macfarlane and Donath (1988). In qualitative reasoning the algebraic signs of the relationships between variables and parameters are considered, as is also done in the present work. However, these theories employ *passive* approaches to observe the effects of physical processes, but no decisions are made. Design requires an *active* approach where, in addition to predicting process and functionality, decisions are required.

This paper proposes that *the creation of innovative designs of physical significance, concerning geometric and material properties, requires reasoning from first principles.* Previous work emphasizing the importance of this type of reasoning in structural design can be found in PROMPT (Murthy and Addanki, 1987). PROMPT is significantly different in methodology from the present work. PROMPT uses human pre-compiled knowledge to create design modification operators. Our approach, known as 1stPRINCE (FIRST PRINciple Computational Evaluator), provides a mechanism to automate the process of creating operators which are derivable from first principle knowledge. 1stPRINCE utilizes monotonicity analysis on fundamental structural equations to traverse the solution space and discover parametric designs demonstrating optimal behavior. In the PROMPT publication, an example of the design of a beam under torsion is discussed. This example is well understood by engineers and serves as a good example for the application of our methodology. It will thus be utilized within this paper and used for comparison with the PROMPT approach.

Design is a goal-oriented process (Radford and Gero, 1985) with a design objective and constraints on the physical parameters which bound the solution and model functional relations between variables of physical significance (throughout this paper, 'function' refers to mathematical function and not process function, unless otherwise stated). Because numerical design optimization requires well-formulated numerical problem specifications, it is typically only useful in the detailing stages of the design process. Although rational design involves optimizing behavior, a human designer may reason qualitatively rather than numerically to observe the entire design space symbolically and make design improvements. 1stPRINCE first performs qualitative optimization of a design problem by means of monotonicity analysis. If all constraints cannot be met from a monotonicity analysis, additional first principle information is utilized to make fundamental design modifications. This reasoning considers the significance of mathematical quantities such as the integral and utilizes heuristics on how the design space can be divided into separate regions. As physical constraints are determined to be active (i.e., relevant in their limit at optimality), they are backsubstituted into the objective function to obtain a solution with minimal degrees of freedom at the optimum. This important functional information for structures is often expressed in terms of material properties, which are variables in this system, and allows a designer (or an expert system with access to a library of materials) to choose a material which will allow the objective function to be minimized and obtain the proper dimensions to reach the design goal.

Reasoning via this approach makes possible the conceptual design of structures by utilizing first principles. Application of 1stPRINCE to the design of a shaft under torsion yields interesting results. From knowledge (fundamental equations) only of a solid round shaft, two optimizing design modifications are obtained. A hollow tube is created with a large radius constrained by either minimum thickness or by buckling. A composite rod is also discovered with a cylinder of inside material bonded to a hollow tube of outside material. Here limitations on the optimal design occur in the interaction between the material properties of the two sections. If certain derived conditions hold and one material is heavier, stiffer, and has a higher limit on strength than the other, then it is preferred to be the outside material with as small a thickness as practical. All of these results are obtained directly from application of our design theory with no pre-compiled knowledge about mechanical behavior. *1stPRINCE makes decisions which are optimally directed*; *it does not haphazardly try new combinations.*

Following the nomenclature section, the design of a shaft under a torsion load, our design example, is described along with a summary of past work concerning this example. 1stPRINCE is then described and applied to the torsion problem.

2. Nomenclature

The following nomenclature is utilized in this paper

A, dA	cross-sectional area of beam region and differential
g_j	inequality constraint, j
G	shear modulus of elasticity
h_i	equality constraint, i
$K_j(G)_{(j=1,2)}$	functional relationship between material properties with shear modulus ($j=1$ for density, $j=2$ for yield stress)
L, dz	length of beam and differential
r, dr	radius of circular beam and differential
t_{min}	minimum allowable thickness of material
T	torque acting on beam, a system variable
T_{min}	applied torque acting on beam, a parameter
X_j	dimensions and material properties separated for distinct regions j
W	weight of beam
W_{max}	maximum allowable weight of beam, a parameter
ϕ	angle of rotation of beam
ρ	material density
τ	shear stress
τ_y	yield stress

3. A design example: design of a shaft under torsion

A typical structural design can be seen in Figure 1. A designer must design a shaft of minimum weight (constrained to be less than W_{max}) subject to a torsion load (T) such that the maximum stress remains under yield stress (τ_y) possibly divided by some factor of safety.

The designer may reason as follows

A solid, round bar resists torsion well and is commercially available in almost any size. Choose a solid round bar and determine an appropriate radius to meet the stress constraint. Once the stress constraint is met the weight of the bar should be calculated and the weight constraint checked. If all constraints are met the design is satisfactory and consideration of a more weight-efficient design can be considered. If, however, the bar is too heavy then the design must be modified.

Figure 2(a) shows the stress distribution across a circular cross-section of the bar. Note that the stress in the center is zero and increases linearly as the radius increases. If a small radius r_1 of material is removed from the center portion, the volume (and thus the weight) can be reduced without significantly affecting the stress at the outside radius r_2 (Figure 2(b)).

The expert system PROMPT (Murthy and Addanki, 1987) represents a computational methodology which appears to reason through a problem as just described. PROMPT uses a multilevel approach to design. The first level utilizes typical AI technology. Hierarchical decompositions and atomic components are stored in Prototypes. A Prototype is a data structure containing basic equations, relations and definitions (Addanki and Davis, 1985). A more recent publication describes the use of the Graph of Models data structure in which the nodes of the graph represent the Prototype model and edges represent the assumptions which must be satisfied in order to

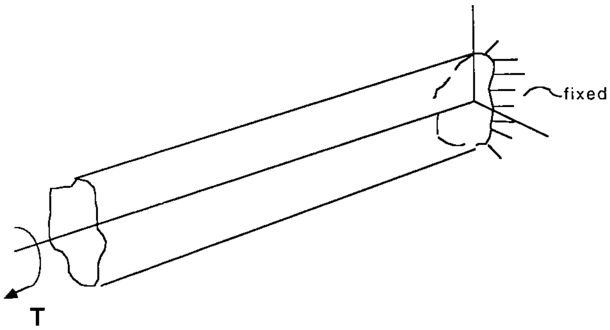

FIGURE 1. General beam under a torsion load T

traverse between nodes (Murthy and Addanki, 1987). By utilizing constraint propagation, a design is modified until all equality constraints are met. If the first level fails to produce a good design the second level is employed. Here 'modification operators' are described as pre-compiled, frequently utilized design changes which are incorporated into the design to attempt to meet constraints. Typical operators are the redistribution of mass and the alteration of shape.

Within PROMPT, rules and heuristics are derived as modification operators from first principles via a human compiler, but the system which is automated does not reason *directly* from first principles. Modification operators are able to make structural modifications such as rounding a corner to relieve stress intensity and are efficient in situations which lie within the domain encompassed by the heuristics. However, modifications not specified by the heuristics cannot take place; much information which can be derived from engineering equations is lost by the human transfer of knowledge from functional form to heuristic form using pre-compiled operators. PROMPT, then, works from fundamental knowledge, but not truly from first principles. The system is capable of deriving modifications in the shape of

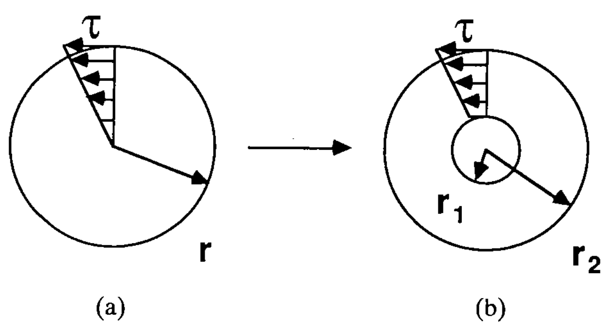

FIGURE 2. Shear stress distribution of a circular beam under torsion load (solid (a) hollow (b)). Stress increases linearly from the center toward the outer radius

prototypes, but does not utilize powerful information which can be derived from the functional relationships between parameters. Further, the system does not reason with inequality constraints in order to optimize the design. It is this information which allows an engineer to decide which solution is preferable when there are several design options.

4. 1stPRINCE

Agogino and Almgren (1987a, b) argue that an ideal design methodology for mechanical systems should be able to move between the qualitative, functional, and numerical levels of human reasoning (Figure 3). The methodology presented in this paper incorporates all three levels of reasoning. Monotonicity analysis (qualitative reasoning) observes the algebraic first derivative of parameters to make qualitative decisions to direct globally optimal behavior or detect flaws in the problem formalization. A design expert system often looks for any feasible solution to a problem (satisficing). With the aid of monotonicity analysis the *best* solution to the same problem based on qualitative information may be found. Mathematical functional information is used as deemed important from the monotonicity analysis. If constraints cannot be met or a designer wishes to investigate alternative designs, a mathematical reasoning module breaks integrals into

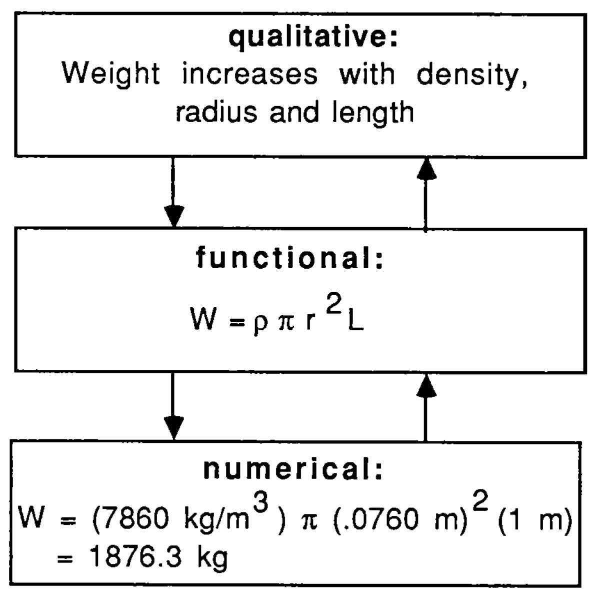

FIGURE 3. An ideal methodology for mechanical design should incorporate the qualitative, functional, and numerical levels and allow for the transition between the three levels. The specific example concerns the weight of a beam of circular cross-section

smaller ranges creating novel design primitives and allowing for innovative designs. All design recommendations are presented in symbolic form. These algebraic solutions to the design problem can be solved numerically to obtain specific solutions and quantitative comparisons between solutions.

4.1 HEURISTIC SEARCH

In the proposed architecture, an initial primitive must be selected by the user or some object-oriented expert system (e.g., OMDesign: Agogino and Guha, 1987). This selection should be based on heuristic information describing which basic structures are best suited for different loadings. Within individual primitive frames is information on the structure such as fundamental equations of deformation, stress, and volume. Also important constraints on the primitive can be found in the frame slots. This selection gives an initial start for the system to perform an analysis and innovate new designs.

4.2 MONOTONICITY ANALYSIS: QUALITATIVE REASONING

Design is a goal-oriented process with objectives and constraints which can often be formulated as a series of mathematical equations. Numerous numerical techniques exist (Vanderplaats, 1984; Haug and Arora, 1979; Reklaitis *et al.*, 1983) to solve optimization problems described in this manner, however a numerical solution loses generalization and important functional information found in the relationships between parameters. Monotonicity analysis is a symbolic approach to non-linear optimization problems which utilizes information at the qualitative level to determine globally optimum search directions. Originally developed by Papalambros and Wilde (1979) and automated symbolically with AI technology by Choy and Agogino (1986), the analysis utilizes qualitative first derivative information (i.e., the algebraic sign of the direction of change of a variable) to determine which constraints will be active or inactive for a possible solution to the optimization problem. The analysis verifies that a problem formulation is properly bounded and occasionally can identify the global optimum to a problem directly without further analysis.

First-generation expert system techniques lose important functional information (similar to that which numerical solutions lose) by compiling general knowledge into heuristics. Monotonicity analysis, by itself, also loses the same information; however, by backsubstituting the mathematical functional information into the objective function, more powerful solutions can be derived. Monotonicity analysis is described below and functional backsubstitution is described later.

Definitions

a. The *monotonicity* of a continuously differentiable function $f(x)$ with respect to variable x_k is the algebraic sign of $\partial f/\partial x_k$. (Note: the problem need not be formulated with equations; only the relative increase or decrease of one variable with changes in another need be specified. Thus the monotonicity concept can be applied equally well to non-continuous functions.)
b. A constraint $g_i(x) \leq (\geq)\ 0$ is *active* at x_0 if $g_i(x_0) = 0$. A constraint $g_i(x) \leq (\geq) 0$ is *inactive* at x_0 if $g_i(x_0) < (>) 0$.
c. A positive variable x_k is said to be *bounded above* by a constraint $g_i(x) \leq 0$ if it achieves its maximum value at strict equality, i.e. when the constraint is active. A positive variable x_k is *bounded below* by $g_i(x) \leq 0$ if it achieves its minimum value at strict equality.

Figure 4(a) shows a function, $f(x)$, which has an interior minimum (zero first derivative and positive second derivative) within the feasible range between $x = a$ and $x = b$. In Figure 4(b), there is no zero first derivative and because the function is monotonically increasing in x the minimum value of $f(x)$ occurs at the lower boundary. Thus the constraint stating that $x \geq a$ becomes *active*, implying that $x = a$ at optimality ($x \geqq a$). (Note that the symbols '$\geqq$' and '$\leqq$' will be used to designate active inequality constraints or relevant directed equality constraints.)

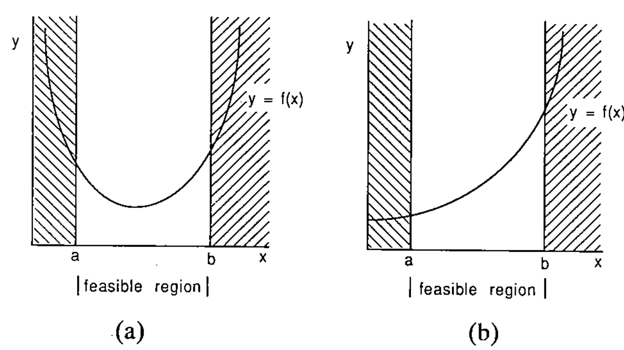

FIGURE 4. The minimum of the function in (a) occurs within the interior of the region. In (b) there is no zero first derivative and the absolute minimum of the feasible range occurs at the lower boundary implying the inequality constraint on the lower bound is active

Monotonicity analysis identifies which constraints would be propagated in order to optimize the objective, but does not activate inequality constraints that would be suboptimal. It can thus be viewed as an optimizing formalism of the 'least commitment' principle in the AI literature (Stefik, 1981). However, unlike conventional constraint propagation which just works with equality constraints, monotonicity analysis works with both equality and inequality constraints.

Three rules of monotonicity analysis define well-constrained optimization problems (Papalambros and Wilde, 1979; Papalambros, 1982) without overconstrained cases (Wilde, 1985).

Rule One: If the objective function is monotonic with respect to (w.r.t.) a variable, then there exists at least one active constraint which bounds the variable in the direction opposite of the objective.

Rule Two: If a variable is not contained in the objective function then it must be either bounded from both above and below by active constraints or not actively bounded at all (i.e., in the latter case, any constraints monotonic w.r.t. that variable must be inactive or irrelevant).

Rule Three (The Maximum Activity Principle): The number of non-redundant active constraints cannot exceed the total number of variables.

Utilizing these three rules, constraints which are unconditionally active or inactive may be determined. The implication of active constraints is that a design variable is being driven toward an upper or lower limit. The implementation of these rules has been automated by Choy and Agogino (1986) in the program SYMON (SYmbolic MONotonicity analyzer) in the Franz LISP (Fateman, 1982) and VAXIMA languages (Rand, 1984). The SYMON user interface allows the problem to be displayed in tabular form utilizing monotonicity tables, as will be done in this paper. For ease of presentation, rules one and two can be formulated into a heuristic which will be used to balance the tables in this paper.

Heuristic: A column of a variable in a table must have at least one '+' and '−', or else nothing. The objective function must be relevant.

An equality constraint is always active, but it is not always significant (*relevant*) in bounding an optimization problem. In other words, irrelevant constraints can be left out without affecting the final solution. In optimization theory, the corresponding Lagrange multipliers would be zero.

4.3 INTRODUCTORY EXAMPLE

A simple example of monotonicity analysis can be found in Table 1 (a). This example will serve as an introduction to readers unfamiliar with the theory, and also has parallels to the more detailed design problem to be presented later in the paper. In the table, equality constraint i is specified by the symbol 'h_i' and inequality constraint j as 'g_j'. Since inequality constraints are bounded, the direction of that bound is given for each variable; the directions of monotonicity of equality constraints are not known before an analysis so question marks are put in the appropriate columns. At this point the activity of both equality

TABLE 1. Simplified formulation of minimization problem to determine the minimum weight of a solid rod, given constant T and L. In (a) an unresolved monotonicity table is given. (b) and (c) Describe two solutions (radius-constrained case and stress-constrained case, respectively) which balance the table, implying a well-formulated and well-constrained problem. Shaded areas signify inactive and irrelevant constraints

	r	τ
obj: weight $\pi L \rho r^2$	+	
g_1 $r \geq r_{min}$	-	
g_2 $\tau \leq \tau_y$		+
h_1 $\tau = \frac{2T}{\pi r^3}$	?	?

unresolved table

(1a)

g_1 active	r	τ
obj: weight $\pi L \rho r^2$	+	
g_1 $r \geq r_{min}$	-	
g_2 $\tau \leq \tau_y$ (shaded)		+
h_1 $\tau = \frac{2T}{\pi r^3}$ (shaded)	?	?

case 1:
radius constrained

(1b)

g_2 active h_1 relevant	r	τ
obj: weight $\pi L \rho r^2$	+	
g_1 $r \geq r_{min}$ (shaded)	-	
g_2 $\tau \leqq \tau_y$		+
h_1 $\tau \geqq \frac{2T}{\pi r^3}$	-	-

case 2:
stress constrained

(1c)

and inequality constraints is not known since the monotonicity analysis has not been performed.

The introductory example problem is to minimize the weight ($\pi L\rho r^2$) of a solid, round bar of given length (L), mass density (ρ), and torque (T). Two inequality constraints ($g_{1,2}$) place a lower bound on the radius by $r_{\min}$ and an upper bound on the shear stress (τ) by τ_y (yield stress), respectively. Equality constraint h_1 gives the relation between shear stress and radius. Actually, an equality constraint can be written as two inequality constraints, only one of which can be active. Thus constraint h_1 could be written as $\tau \geqq (2T)/(\pi r^3)$ and $\tau \leqq (2T)/(\pi r^3)$. Utilizing our heuristic, constraint g_1 may be active to balance the objective function (Table 1(b)). The table would be properly bounded at this point with constraints g_2 and h_1 eliminated. An alternative solution can be found by setting g_2 active, eliminating g_1, and using h_1 to again balance the table by using the equivalent inequality to the constraint where $\tau \geqq (2T)/(\pi r^3)$ (Table 1(c)). Thus there are two possible solutions to this problem, one in which the radius is set at its minimum value and the other where the radius is determined by the shear stress constraint. Both solutions are constraint-bound. Further details of the execution of the analysis can be found in Michelena and Agogino (1988).

4.4 MATHEMATICAL REASONING

Monotonicity analysis can give the candidate cases of well-constrained solutions to an optimization problem. An intelligent design tool should prune these cases to bounded and feasible solutions. As constraints cannot be met, decisions must be made to modify the design from a higher level of reasoning. As modifications in problem formulation take place, paths of the design space which have already been searched should be ignored. One approach to modifying designs is derived from the mathematical quantities which formulate the initial problem.

In particular, the integral appears in numerous fundamental equations of structural mechanics. A continuous integral can be split into continuous integrals on smaller intervals

$$\int_{a_0}^{a_1} f(x)\,\mathrm{d}x = \int_{a_0}^{b} f(x)\,\mathrm{d}x + \int_{b}^{a_1} f(x)\,\mathrm{d}x, \quad \text{etc.,} \quad (1)$$

where $a_0 \leq b \leq a_1$.

Often, in structural engineering, material properties and geometry are constant within a region and an integral is replaced with a summation

$$\int_{a_0}^{a_1} f(x)\,\mathrm{d}x = \sum_{i=1}^{n} f(x_i)\,\Delta x_i, \qquad n = 1, 2, 3, \ldots \quad (2)$$

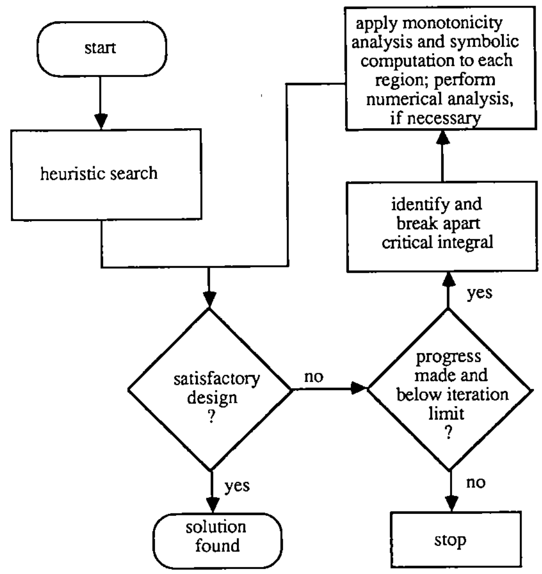

FIGURE 5. 1stPRINCE's method to discover innovative designs by integrating heuristic and first principle reasoning

By breaking an integral into a summation, a region is broken into separate and distinct regions with homogeneous properties within a region, but not across regions. (Continuity must be considered at region boundaries.) Since a continuous integral is a summation over infinitely small divisions, it follows that small but discrete regions can be used to represent the integral formulation in mechanical structures problems. Finite element methods use fine meshes to *numerically* approximate the integral for structural analysis. Our design approach is to first search for strong qualitative trends *symbolically* using a minimal number of regions before resorting to numerical computation (Figure 5).

4.5 INTEGRAL DIVISION ALGORITHM

1. Initially break the integral into one region assuming the properties of the region are constant throughout (i.e., assume constant mass density, shear modulus of elasticity, etc. in the region).

2. Utilize monotonicity analysis on the n-region formulated problem (on first pass $n = 1$) to attempt to find possible optimal solutions. If all constraints are met a solution to the problem has been found.

3. If constraints are not all met, divide each region into two separate regions (e.g., the one-region space is divided into two regions).

4. Go to Step 2 and continue until a desired design or the iteration limit is reached.

After the first division there will be two regions, then four, eight, etc. As the problem grows in space, as long as progress is made, the amount of time it takes to analyze the possible solutions grows combinatorially. Thus it is important to limit the number of iterations to a reasonable number and to keep histories of each analysis so that previous analyses are not repeated. The approach of starting with few regions and breaking each sub-region apart is efficient; if a search directs the region to zero mass then all future iterations can ignore that region. Thus after the second iteration if one of the regions is found to go to zero then the next iteration will find only two and not four regions to analyze. In addition most physical structures are designed to be as simple as possible in order to reduce manufacturing, maintenance and analysis costs; thus fewer regions instead of many regions may be preferred in a good design. Lastly, only a few regions may be all that is necessary to show qualitative trends toward the directions of improvement for the human designer interacting with the system. It is for these reasons that regions are divided into two rather than more regions at each stage.

4.5 CONSTITUTIVE RELATIONS

Constitutive relations imply that certain physical assumptions must be met in order to make a design valid. In addition to constraints on upper and lower bounds of variables, these constitutive relations add important constraints to the problem formulation.

Additional interrelationships that must be considered are those between material properties. Material properties are not independent. This is represented by the influence diagram (Rege and Agogino, 1988) in Figure 6(a). A node with an arrow leaving it states that the node contains information which influences the node to which the arrow leads. If

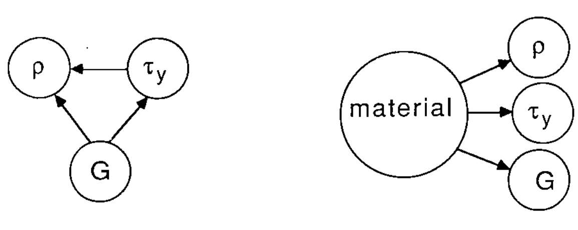

FIGURE 6. Influence diagram of the relationship between variables. In (a) shear modulus influences mass density and yield stress. In (b) a given material defines values of all material properties

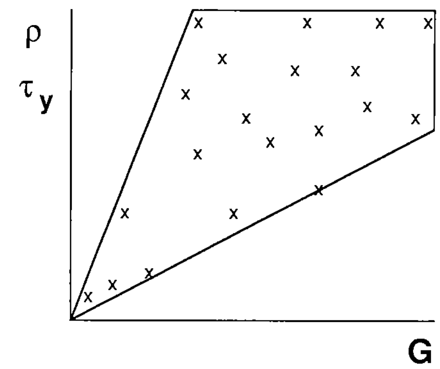

FIGURE 7. Qualitative plot of shear stress and mass density vs. shear modulus for various materials. Note monotonically increasing trend

an arrow joins two nodes, this implies that the two nodes are not (probabilistically or deterministically) independent. The diagram in Figure 6(a) states that the elastic shear modulus, G, influences material density, ρ, and yield stress, τ_y. Of course, this diagram could have been stated in reverse (i.e., ρ influences G and τ_y). If, however, a specific material is given, then specific material properties result, within statistical variations (Figure 6(b)).

Referring to Figure 6(a) again, a general theoretical influence between all material properties (specifically shear modulus versus material density and yield stress) cannot be derived for all materials (although some relationships can be determined like that of shear modulus, Poisson's ratio, and Young's modulus). However, a plot of G vs ρ or τ_y for arbitrary materials (Figure 7) demonstrates a monotonic positive trend; overall, as G increases so do ρ and τ_y. This implies that there is no known ideal material with infinite rigidity, zero mass density, and infinite yield stress. Within the bounded area of Figure 7 a local trend can occur which has negative or non-changing monotonicity, but for this application of monotonicity analysis the global trend is utilized as a positive overall monotonicity. Thus the equality constraints

$$\rho = K_1(G^+) \quad \text{and} \quad \tau_y = K_2(G^+) \qquad (3)$$

(ρ and τ_y are monotonically increasing functions of G) are employed to represent this relationship. Thus material properties are considered as variables in this theory; beside determining geometry, the method determines material property requirements.

As mentioned, constitutive relations also lead to inequality constraints which help bound the design space. For example, practical material properties must have values greater than or equal to zero (in theory, Poisson's ratio could be negative). An additional constitutive relation would state that the beam must have positive volume; it cannot be massless.

In addition to constitutive relations, fundamental engineering assumptions need to be considered, such as boundary conditions. An example of an assumed boundary condition gives zero slope and deflection at the clamped end of a beam.

4.6 FUNCTIONAL RELATIONSHIPS (BACKSUBSTITUTION USING COMPUTER ALGEBRA)

A qualitative analysis of the design problem is completed utilizing monotonicity analysis. Certain valid, unique solution domains occur via active constraints. Much important information can be found from the relationships of the variables within those active constraints. A functional analysis can then be performed on each of the cases by backsubstituting the known, active information into the objective function. When an inequality constraint is deemed active, that constraint is at its bound and the values of the variables are known to equal the bounded values. The constraint can then be assumed to be a relevant equality constraint within the given solution.

Referring again to the introductory example of Table 1 (the simplified problem of a round beam under torsion), both solutions lead to final relative weight equations after backsubstitution. In Table 1(b), the constraint on the outside radius is active and backsubstitution leads to an equation of weight as

$$W = \pi L \rho r_{\min}^2. \tag{4}$$

Table 1(c) requires the shear stress to be at yield for a solution of

$$W = \pi L \rho \left(\frac{2T}{\pi \tau_y}\right)^{2/3} \tag{5}$$

As active constraints are backsubstituted into the objective function, hidden monotonicities can be revealed (Agogino and Almgren, 1987b). Before backsubstitution this problem appears to be well-formulated and well-bounded. As the substitution takes place, certain variables can cancel out of the problem and hidden monotonicities may become apparent, requiring additional constraints to become active to satisfy the rules of monotonicity once again. It is necessary to recheck monotonicities during backsubstitution and add new constraints to the active list as required to satisfy the rules of monotonicity analysis. For a well-bounded problem if all constraints are backsubstituted into the objective function, then the objective function must be either non-monotonic or constant. All active constraints must be used during backsubstitution since each constraint contains important information about the optimal solution.

After the above process takes place, the objective function contains powerful information about the optimal solution. Regions which benefit most from certain ranges of variable values can be determined. If one variable is raised to a higher power than a different one, the objective function will benefit more by varying that parameter in the optimizing direction. Specific examples will be demonstrated in the next section as 1stPRINCE is applied to the 'Torsional Beam' problem.

5. Solution to torsional beam problem

5.1 THE FIRST PASS

In this section the 1stPRINCE methodology described in the previous section is applied to the design problem of a beam resisting a torsional load. Elementary equations of the deformation and weight of a beam under torsion are:

$$\tau = \frac{G\varphi r}{L} \quad \text{(shear stress)},$$

$$\varphi = \frac{TL}{\int r^2 G \, \mathrm{d}A} \quad \text{(angle of rotation)}, \tag{6}$$

$$W = \iint \rho \, \mathrm{d}A \, \mathrm{d}z \quad \text{(weight)},$$

where G is the shear modulus, ρ is the mass density, A is the cross-sectional area, L is the beam length (of differential dz), T is the applied torque, and r is the radius (employing polar coordinates).

An assumption that there is no variation in variables along the length of the beam is now made in order to reduce the complexity of the presentation. Initially a solid, circular shaft ($A = \pi r_1^2$) is employed. This is the initial choice of the designer in the previous discussion of this problem and would be a good selection by an expert system from a library of basic beams. Equations (6) can be reduced (with the first reduction due to length and the second due to area) to

$$\tau = \frac{G\varphi r}{L} = \frac{G\varphi r_1}{L},$$

$$\varphi = \frac{TL}{\int_0^{r_1} 2\pi r^3 G \, \mathrm{d}r} = \frac{2TL}{\pi r_1^4 G} \tag{7}$$

$$W = \int_0^{r_1} 2\pi \rho L r \, \mathrm{d}r = \pi \rho L r_1^2.$$

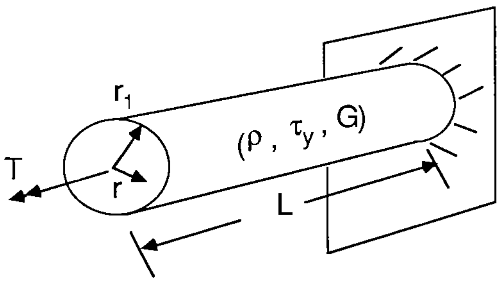

FIGURE 8. Circular, solid beam which is clamped at one end and undergoing a torque at the free end

The reduced formula for shear stress represents its maximum value within the region. In reference to the proposition about dividing the integral into sub-regions, this is the first pass where there is only one unit having constant material properties. The beam now appears as in Figure 8.

This design problem can be formulated as an optimization problem where the weight is minimized subject to various constraints. Equality constraints evolve from constitutive relations and fundamental relations as found in equations (7). Inequality constraints require that the variables be bounded and evolve from constitutive relations and realistic design considerations. Two important constraints require that the stress be below the yield limit (modified by a factor of safety) and the weight be below some maximum value. The problem will exclude the weight constraint for initial analysis and any derived solution will then be checked to verify that the constraint is met.

Keeping L and T constant the optimization problem is summarized in the unresolved monotonicity tabular form in Table 2. Two physically reasonable solutions

TABLE 2. Unresolved monotonicity table representing the objective function and constraints for a beam under torsion during the first iteration

◯	τ	ρ	G	r_1	φ	τ_y	T
objective: $\pi \rho L r_1^2$		+		+			
constraints:							
g_1 $\tau \le \tau_y$	+					-	
g_2 $\rho \ge 0$		-					
g_3 $G \ge 0$			-				
g_4 $r_1 \ge 0$				-			
g_5 $\varphi \le \varphi_{max}$					+		
g_6 $\tau_y \ge 0$						-	
g_7 $T \ge T_{min}$							-
h_1 $\tau = \frac{G \varphi r_1}{L}$	?		?	?	?		
h_2 $\varphi = \frac{2 T L}{\pi G r_1^4}$			?	?	?		?
h_3 $\rho = K_1(G)$		?	?				
h_4 $\tau_y = K_2(G)$			?			?	

can be found in Tables 3 and 4. Inactive or irrelevant constraints are shaded in the tables. Case 1 (Table 3) requires constraints g_1, g_2 and g_7 to be active and h_1, h_2 and h_4 to be relevant. The stress must be at yield and the density at zero. There is no material with zero density, however this statement implies that the optimum practical solution occurs with as small a density as possible; substitution of $\rho = 0$ into the objective gives $W = 0$. Of course this solution obviously will not meet the inactive constraints, but the methodology allows a computer to determine this information and continue to more interesting cases.

Case 2 (Table 4) is a more realistic situation in which constraints g_1, g_7, h_1, h_2, h_3 and h_4 are active and relevant, implying that the stress be at yield and the relationship between density and shear modulus be relevant; shear modulus is constraining density from going to a bound and visa versa. Also the total torque which the design can resist is the applied torque.

Now active inequality and relevant equality constraints can be backsubstituted into the objective function as previously discussed. Utilizing the following six equations:

$$
\begin{aligned}
\tau &= \tau_y, \\
\tau &= \frac{G\varphi r_1}{L}, \\
\varphi &= \frac{2TL}{\pi G r_1^4}, \\
T &= T_{min}, \\
\rho &= K_1(G), \\
\tau_y &= K_2(G),
\end{aligned}
\tag{8}
$$

TABLE 3. Case 1: Mass density approaches zero

Case 1 ◯	τ	ρ	G	r_1	φ	τ_y	T
objective: $\pi \rho L r_1^2$		+		+			
constraints:							
g_1 $\tau \leqq \tau_y$	+					-	
g_2 $\rho \geqq 0$		-					
g_3 $G > 0$			-				
g_4 $r_1 > 0$				-			
g_5 $\varphi < \varphi_{max}$					+		
g_6 $\tau_y > 0$						-	
g_7 $T \geqq T_{min}$							-
h_1 $\tau \geqq \frac{G \varphi r_1}{L}$	-		+	+	+		
h_2 $\varphi \geqq \frac{2TL}{\pi G r_1^4}$			–	–	–		+
h_3 $\rho = K_1(G)$		?	?				
h_4 $\tau_y \leqq K_2(G)$			-			+	

TABLE 4. Case 2: Constraint h_3 keeps density from approaching zero

Case 2	τ	ρ	G	r_1	φ	τ_y	T
objective: $\pi \rho L r_1^2$		+		+			
constraints:							
g_1 $\tau \leq \tau_y$	+					-	
g_2 $\rho > 0$		-					
g_3 $G > 0$			-				
g_4 $r_1 > 0$				-			
g_5 $\varphi < \varphi_{max}$					+		
g_6 $\tau_y > 0$						-	
g_7 $T \geq T_{min}$							-
h_1 $\tau \geq \frac{G \varphi r_1}{L}$	-		+	+	+		
h_2 $\varphi \geq \frac{2 T L}{\pi G r_1^4}$			−	−	−		+
h_3 $\rho \geq K_1(G)$		−	+				
h_4 $\tau_y \leq K_2(G)$			-			+	

the objective can be found to be

$$W = \pi L \rho \left(\frac{2T_{\min}}{\pi \tau_y} \right)^{2/3}, \tag{9}$$

or

$$W = \pi L K_1(G) \left(\frac{2T_{\min}}{\pi K_2(G)} \right)^{2/3}. \tag{10}$$

The weight is now a function of density and yield stress, but more specifically only a function of shear modulus (i.e., a given material) for a given length and applied torque. As expected, finding a material with low density and high yield stress will give an optimally behaved design.

The beam dimensions are also determined via backsubstitution and the outside radius as a function of material and torque becomes

$$r_1 = \left(\frac{2T_{\min}}{\pi \tau_y} \right)^{1/3} = \left(\frac{2T_{\min}}{\pi K_2(G)} \right)^{1/3}. \tag{11}$$

The size of the beam for minimum weight is now a function of the chosen material and applied torque *only*.

5.2 THE SECOND PASS

If, given the beam dimensions and material properties, the weight is calculated and found to exceed the weight constraint or no material properties

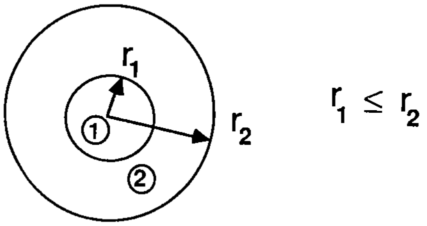

FIGURE 9. Division of single region into two separate and distinct regions from the second iteration

can satisfy the constraint-bound case, the second iteration should take place to modify the original design. Even if all constraints are met this second iteration may be desired to allow the designer opportunities to create a more elegant design. The original objective function,

$$W = 2\pi L \int_0^{r_2} \rho r \, dr, \tag{7c}$$

appears in integral form. On second pass, the integral in the objective is divided into two regions, the first from 0 to r_1, the second from r_1 to r_2:

$$2\pi L \int_0^{r_2} \rho r \, dr = 2\pi L \int_0^{r_1} \rho_1 r \, dr + 2\pi L \int_{r_1}^{r_2} \rho_2 r \, dr. \tag{12}$$

This division is represented in Figure 9. The actual value of r_1, in terms of r_2, is undetermined and will be found on the basis of functional information. There now exist two separate regions, each of distinct material properties and fundamental relationships. The new problem is formulated in Table 5. The objective function becomes the minimization of the addition of the weights of the two regions. Each variable now needs to be bounded. Geometric considerations require that the outside radius be greater than the inside radius by some minimum thickness, t_{min} (constraint g_{11}), and the total shaft must resist at least the applied torque, T_{min} (constraint g_{13}).

TABLE 5. Objective function and constraints for second iteration

Objective: $W_1 + W_2$			
Equality constraints		Inequality constraints	
h_1	$W_1 = \pi L \rho_1 r_1^2$	g_1	$\tau_1 \le \tau_{y1}$
h_2	$W_2 = \pi L \rho_2 (r_2^2 - r_1^2)$	g_2	$\tau_2 \le \tau_{y2}$
h_3	$\varphi_1 = \varphi_2$	g_3	$\rho_1 \ge 0$
h_4	$\varphi_1 = (2T_1 L)/[\pi G_1 r_1^4]$	g_4	$\rho_2 \ge 0$
h_5	$\varphi_2 = (2T_2 L)/[\pi G_2 (r_2^4 - r_1^4)]$	g_5	$G_1 \ge 0$
h_6	$\tau_1 = G_1 \varphi_1 r_1 / L$	g_6	$G_2 \ge 0$
h_7	$\tau_2 = G_2 \varphi_2 r_2 / L$	g_7	$W_1 \ge 0$
h_8	$\rho_1 = K_1(G_1)$	g_8	$W_2 \ge 0$
h_9	$\rho_2 = K_1(G_2)$	g_9	$r_1 > 0$
h_{10}	$\tau_{y1} = K_2(G_1)$	g_{10}	$r_2 > 0$
h_{11}	$\tau_{y2} = K_2(G_2)$	g_{11}	$r_2 \ge r_1 + t_{min}$
		g_{12}	$r_2 \le r_{max}$
		g_{13}	$T_1 + T_2 \ge T_{min}$
		g_{14}	$\varphi_2 \le \varphi_{max}$
		g_{15}	$\tau_{y1} \ge 0$
		g_{16}	$\tau_{y2} \ge 0$

One additional requirement is that of continuity across the boundary between the two regions (constraint h_3). The assumption made that the two angles be equivalent implies that the bond is perfectly rigid, an assumption often made in linear beam theory. An analogous situation is the assumption of a beam clamped into a wall where the slope of the beam at the wall is assumed to be zero. Here one region can be thought of as clamped to the other and thus the angles must be equivalent. This type of fundamental engineering assumption must be included within the knowledge base.

This new problem can again be modelled in a monotonicity table. Since the use of a monotonicity table is only to allow the user to observe the analysis, certain constraints which will be unconditionally inactive can be deleted. Since the weights, densities, and yield stresses are all dependent on shear moduli, only the constraints requiring $G_i \ge 0$ ($g_{5,6}$) are necessary while positivity conditions $g_{3,4,7,8,15,16}$ will be ignored (a default assumption that must be verified by the final solution). In addition, $g_{9,10}$ will be ignored since the case where the inside radius approaches zero reduces the problem to that of the first iteration and that of the outside radius approaching zero implies a beam of zero mass which has already been eliminated. This reduced, but unresolved, monotonicity table is shown in Table 6.

Further observations can be made from the table using the first rule of monotonicity. Since the objective function is relevant, either the weights themselves (actually the shear moduli) must approach zero (constraints $g_{5,6}$, but not both of them!) or the weight equations must be relevant ($h_{1,2}$). If the weights don't go to zero then the equality constraint for angular deflection (h_3) must be relevant and one or both of the maximum stresses in a region must be at yield ($g_{1,2}$). The outside material cannot vanish or the problem would again reduce to the first iteration case (g_6 is unconditionally inactive). Angle definitions ($h_{4,5}$) must always be relevant and balance the unconditionally active torque constraint (g_{13}). Therefore the optimal solution will always design for the applied torque. Through observations such as these, the search space is being pruned to ignore redundant cases.

TABLE 6. Unresolved monotonicity table for relevant constraints for second iteration

① 2	W_1	W_2	τ_1	τ_2	ρ_1	ρ_2	G_1	G_2	r_1	r_2	φ_1	φ_2	T_1	T_2	τ_{y1}	τ_{y2}
obj: $W_1 + W_2$	+	+														
h_1 $W_1 = \pi L \rho_1 r_1^2$	?				?				?							
h_2 $W_2 = \pi L \rho_2 (r_2^2 - r_1^2)$		?				?			?	?						
h_3 $\varphi_1 = \varphi_2$											?	?				
h_4 $\varphi_1 = (2T_1 L)/[\pi G_1 r_1^4]$							?		?		?		?			
h_5 $\varphi_2 = (2T_2 L)/[\pi G_2 (r_2^4 - r_1^4)]$								?	?	?		?		?		
h_6 $\tau_1 = (G_1 \varphi_1 r_1)/L$			?				?		?		?					
h_7 $\tau_2 = (G_2 \varphi_2 r_2)/L$				?				?		?		?				
h_8 $\rho_1 = K_1(G_1)$					?		?									
h_9 $\rho_2 = K_1(G_2)$						?		?								
h_{10} $\tau_{y1} = K_2(G_1)$							?								?	
h_{11} $\tau_{y2} = K_2(G_2)$								?								?
g_1 $\tau_1 \le \tau_{y1}$			+												-	
g_2 $\tau_2 \le \tau_{y2}$				+												-
g_5 $G_1 \ge 0$							-									
g_6 $G_2 \ge 0$								-								
g_{11} $r_2 \ge r_1 + t_{min}$									+	-						
g_{12} $r_2 \le r_{max}$										+						
g_{13} $T_1 + T_2 \ge T_{min}$													-	-		
g_{14} $\varphi_2 \le \varphi_{max}$												+				

Implied from the above discussion, only two unique physically valid structural solutions to the optimization problem exist: the first (Table 7) is the solution of the hollow tube and the second, a composite rod (Table 8).

5.3 CASE 1: HOLLOW TUBE

Table 7 shows g_5 active, letting $G_1 = 0$. Constraint h_8 active gives $\rho_1 = 0$, and g_1 is inactive. Thus there is no stiffness or mass in region 1 and no requirements on the stress, implying no material in that region. The tube is thus hollow. The optimal design further requires that the stress at the outside radius of region 2 be at yield (g_2) and the torque at the applied level (g_{13}) as previously discussed. This information is backsubstituted into the objective. Hidden monotonicities are discovered and additional constraints must be made active in order to place a lower bound on r_2. Two possibilities exist. Constraint g_{11} could be active, stating that the outside thickness is at some minimum and the outside radius can be as large as necessary; otherwise the design becomes constrained by buckling and requires an additional constraint (Timoshenko and Gere, 1961)

$$r_2 - r_1 \ge \left\{\frac{9(1-v^2)^{3/2}}{4\pi^2(1+v)^2}\right\}^{1/5}\left\{\frac{T^2}{G^2(r_1+r_2)}\right\}^{1/5} N, \quad (13)$$

where v is Poisson's ratio which exhibits little variation between materials and is not considered as a variable in the optimization problem, and N is a factor of safety. This equation follows thin shell assumptions $((r_2 + r_1)/(r_2 - r_1) \ge 20)$. The buckling constraint is necessary when considering hollow tubes and is a constraint which must be added when required. Equation (13) leads to complicated symbolic equations and if this constraint is active the solution would best be solved numerically. Since buckling was not