Modern Homotopy Methods in Optimization

L.T. Watson and R.T. Haftka

TR 88-51

November 14, 1988

MODERN HOMOTOPY METHODS IN OPTIMIZATION

Layne. T. Watson
Department of Computer Science
Raphael T. Haftka
Department of Aerospace and Ocean Engineering
Virginia Polytechnic Institute & State University
Blacksburg, VA 24061

Abstract. Probability-one homotopy methods are a class of algorithms for solving nonlinear systems of equations that are accurate, robust, and converge from an arbitrary starting point almost surely. These new techniques have been successfully applied to solve Brouwer fixed point problems, polynomial systems of equations, and discretizations of nonlinear two-point boundary value problems based on shooting, finite differences, collocation, and finite elements. This paper summarizes the theory of globally convergent homotopy algorithms for unconstrained and constrained optimization, and gives some examples of actual application of homotopy techniques to engineering optimization problems.

1. Introduction. Continuation in various forms has been used for a long time in mathematics and engineering, with such names as parameter continuation, incremental loading, displacement incrementation, imbedding, invariant imbedding, continuous Newton, and homotopy. The state-of-the-art of continuation methods was thoroughly surveyed in [1], and more recently in [17]. Recent mathematical developments have led to a whole new class of continuation methods known as *probability-one homotopy algorithms*, which have been successfully applied to solve Brouwer fixed point problems, polynomial systems of equations, and discretizations of nonlinear two-point boundary value problems based on shooting, finite differences, collocation, and finite elements. These new techniques have recently begun to be applied to optimization, and have found significant application in solving some engineering optimization problems.

Homotopy methods are very powerful, robust, accurate, numerically stable, and almost universally applicable, but also often prohibitively expensive. They are particularly suitable for highly nonlinear problems for which initial solution estimates are difficult to obtain. Properly implemented they are indeed *globally convergent*, i.e., converge to a solution from an *arbitrary* starting point. This (costly) global convergence feature is their forte, but also makes them inappropriate for mildly nonlinear problems or problems for which a good initial estimate of the solution is easily obtained.

The objectives of this paper are to summarize the basic theory of globally convergent homotopy methods relevant to optimization, to show how homotopy algorithms may be applied to solve optimization problems, and to give some actual engineering applications. Section 2 gives an intuitive explanation of what is different about the new globally convergent homotopy algorithms, and Section 3 briefly recounts the basic mathematical theory. Section 4 summarizes basic homotopy results for optimization, and makes the connection between nonlinear equations, homotopies, and optimization. Examples of the globally convergent homotopy techniques applied to optimization are given in Sections 5–8.

2. Continuation versus homotopy. Continuation is a well known and established procedure in numerical analysis. The idea is to continuously deform a simple (easy) problem into the given (hard) problem, while solving the family of deformed problems. The solutions to the deformed problems are related, and can be tracked as the deformation proceeds. The function describing the deformation is called a *homotopy map*. Homotopies are a traditional part of topology, and have found significant application in nonlinear functional analysis and differential geometry. Similar ideas, such as incremental loading, are also widely used in engineering.

These traditional continuation algorithms have serious deficiencies, which have been removed by modern homotopy algorithms. The differences, however, are subtle and mathematically deep, and the mathematical proofs of the statements in this article are beyond the scope of the presentation here. To explain the differences between the old and new homotopy techniques, a more detailed discussion is required. Suppose the given problem is to find a root of the nonlinear equation $f(x) = 0$, and that $s(x) = 0$ is a simple version of the given problem with an easily obtainable unique solution x_0. Then a homotopy map could be, e.g.,

$$H(\lambda, x) = \lambda\, f(x) + (1 - \lambda)\, s(x), \quad 0 \leq \lambda \leq 1.$$

The family of problems is $H(\lambda, x) = 0$, $0 \leq \lambda \leq 1$, and the idea would be to track the solutions of $H(\lambda, x) = 0$, starting from $(\lambda, x) = (0, x_0)$, as λ goes from 0 to 1. If everything worked out well, this would lead to a point $(\lambda, x) = (1, \bar{x})$, where $f(\bar{x}) = 0$. The "standard" approach is to start from a point (λ_i, x_i) with $H(\lambda_i, x_i) = 0$, and solve the problem $H(\lambda_i + \Delta\lambda, x) = 0$ for x, with $\Delta\lambda$ being a sufficiently small, fixed, positive number. The bad things that can happen are:

1) The points (λ_i, x_i) may diverge to infinity as $\lambda \to 1$.
2) The problem $H(\lambda_i + \Delta\lambda, x) = 0$ may be singular at its solution, causing numerical instability.
3) There may be no solution of $H(\lambda_i + \Delta\lambda, x) = 0$ near (λ_i, x_i).

The modern approach to homotopy methods is to construct a homotopy map $\rho_a(\lambda, x)$, involving additional parameters in the vector a, such that 1), 2), and 3) never occur or never cause any difficulty. The details of how this is done are given in the next section.

3. Homotopy theory. The theoretical foundation of all probability one globally convergent homotopy methods is given in the following differential geometry theorem:

DEFINITION. *Let E^n denote n-dimensional real Euclidean space, let $U \subset E^m$ and $V \subset E^n$ be open sets, and let $\rho : U \times [0,1) \times V \to E^n$ be a C^2 map. ρ is said to be transversal to zero if the Jacobian matrix $D\rho$ has full rank on $\rho^{-1}(0)$.*

PARAMETRIZED SARD'S THEOREM [4]. *If $\rho(a, \lambda, x)$ is transversal to zero, then for almost all $a \in U$ the map*

$$\rho_a(\lambda, x) = \rho(a, \lambda, x)$$

is also transversal to zero; i.e., with probability one the Jacobian matrix $D\rho_a(\lambda, x)$ has full rank on $\rho_a^{-1}(0)$.

The import of this theorem is that the zero set $\rho_a^{-1}(0)$ consists of smooth, nonintersecting curves in $[0,1) \times V$. These curves are either closed loops, or have endpoints in $\{0\} \times V$ or $\{1\} \times V$, or go to infinity. Another important consequence is that these curves have finite arc length in any compact subset of $[0,1) \times V$. The recipe for constructing a globally convergent homotopy algorithm to solve the nonlinear system of equations

$$F(x) = 0, \tag{1}$$

where $F : E^n \to E^n$ is a C^2 map, is as follows: For an open set $U \subset E^m$ construct a C^2 homotopy map $\rho : U \times [0,1) \times E^n \to E^n$ such that

1) $\rho(a, \lambda, x)$ is transversal to zero,

2) $\rho_a(0, x) = \rho(a, 0, x) = 0$ is trivial to solve and has a unique solution x_0,

3) $\rho_a(1, x) = F(x)$,

4) $\rho_a^{-1}(0)$ is bounded.

Then for almost all $a \in U$ there exists a zero curve γ of ρ_a, along which the Jacobian matrix $D\rho_a$ has rank n, emanating from $(0, x_0)$ and reaching a zero $\bar{x}$ of F at $\lambda = 1$. This zero curve γ does not intersect itself, is disjoint from any other zeros of ρ_a, and has finite arc length in every compact subset of $[0,1) \times E^n$. Furthermore, if $DF(\bar{x})$ is nonsingular, then γ has finite arc length.

The general idea of the algorithm is now apparent: just follow the zero curve γ emanating from $(0, a)$ until a zero $\tilde{x}$ of $F(x)$ is reached (at $\lambda = 1$). Of course it is nontrivial to develop a viable numerical algorithm based on that idea, but at least conceptually, the algorithm for solving the nonlinear system of equations $F(x) = 0$ is clear and simple. The homotopy map (usually, but not always) is

$$\rho_a(\lambda, x) = \lambda F(x) + (1 - \lambda)(x - a), \tag{2}$$

which has the same form as a standard continuation or embedding mapping. However, there are two crucial differences. First, in standard continuation, the embedding parameter λ increases monotonically from 0 to 1 as the trivial problem $x - a = 0$ is continuously deformed to the problem $F(x) = 0$. The present homotopy method permits λ to both increase and decrease along γ with no adverse effect; that is, turning points present no special difficulty. The second important difference is the use of the extraneous parameter a, whose consequence is that there are never any "singular points" which afflict standard continuation methods. The way in which the zero curve γ of ρ_a is followed and the full rank of $D\rho_a$ along γ guarantee this.

In order for property 4) above to hold for the homotopy map in (2), $F(x)$ and $(x - a)$ must be "asymptotically similar" (see Lemma 3 below). This is not the case for *every* $F(x)$, and so frequently other homotopy maps must be used, for example,

$$\rho_a(\lambda, x) = \lambda F(x) + (1 - \lambda) G(x; a), \tag{2a}$$

where $G(x; a)$ is a simple version of $F(x)$. For instance, $G(x; a)$ might be derived by simplifying the physical model used to derive $F(x)$. Also the homotopy map need not be a simple convex combination between $F(x)$ and $G(x; a)$; examples of homotopy maps nonlinear in λ are in [21] and [22].

The scheme just described is known as a probability-one globally convergent homotopy algorithm. The phrase "probability-one" refers to the almost any choice for a, and the "global convergence" refers to the fact that the starting point x_0 need not be anywhere near the solution $\bar{x}$. It should be emphasized that the form of the homotopy map $\rho_a(\lambda, x)$ in (2) is just a special case used here for clarity of exposition. The more general theory can be found in [14, 17, 18], and practical engineering problems requiring a ρ_a nonlinear in λ are in [21] and [22]. Below are some typical theorems for various classes of problems.

The computation of Brouwer fixed points represents one of the first successes for both simplicial [1, 12] and continuous homotopy methods [4, 14]. Brouwer fixed point problems can be very nasty, and often cause locally convergent iterative methods a great deal of difficulty.

THEOREM [14]. *Let $B = \{x \in E^n \mid \|x\|_2 = 1\}$ be the closed unit ball, and $f : B \to B$ a C^2 map. Then for almost all $a \in \text{int } B$ there exists a zero curve γ of*

$$\rho_a(\lambda, x) = \lambda(x - f(x)) + (1 - \lambda)(x - a),$$

along which the Jacobian matrix $D\rho_a(\lambda, x)$ has full rank, emanating from $(0, a)$ and reaching a fixed point $\bar{x}$ of f at $\lambda = 1$. Furthermore, γ has finite arc length if $I - Df(\bar{x})$ is nonsingular.

Typically a mathematical problem (such as a partial differential equation) results in a finite dimensional nonlinear system of equations, and what is desired are conditions on the original problem, not on the final discretized problem. Thus the results in this section are used to derive, working backwards, useful conditions on the original problem, whatever it might be. The following four lemmas, which follow from the results of [4], are used for that purpose.

LEMMA 1. *Let $g : E^p \to E^p$ be a C^2 map, $a \in E^p$, and define $\rho_a : [0, 1) \times E^p \to E^p$ by*

$$\rho_a(\lambda, y) = \lambda g(y) + (1 - \lambda)(y - a).$$

Then for almost all $a \in E^p$ there is a zero curve γ of ρ_a emanating from $(0, a)$ along which the Jacobian matrix $D\rho_a(\lambda, y)$ has full rank.

LEMMA 2. *If the zero curve γ in Lemma 1 is bounded, it has an accumulation point $(1, \bar{y})$, where $g(\bar{y}) = 0$. Furthermore, if $Dg(\bar{y})$ is nonsingular, then γ has finite arc length.*

LEMMA 3. *Let $F : E^p \to E^p$ be a C^2 map such that for some $r > 0$, $x\,F(x) \geq 0$ whenever $\|x\| = r$. Then F has a zero in $\{x \in E^p \mid \|x\| \leq r\}$, and for almost all $a \in E^p$, $\|a\| < r$, there is a zero curve γ of*

$$\rho_a(\lambda, x) = \lambda F(x) + (1 - \lambda)(x - a),$$

along which the Jacobian matrix $D\rho_a(\lambda, x)$ has full rank, emanating from $(0, a)$ and reaching a zero $\bar{x}$ of F at $\lambda = 1$. Furthermore, γ has finite arc length if $DF(\bar{x})$ is nonsingular.

Lemma 3 is a special case of the following more general lemma.

LEMMA 4. *Let $F : E^p \to E^p$ be a C^2 map such that for some $r > 0$ and $\tilde{r} > 0$, $F(x)$ and $x - a$ do not point in opposite directions for $\|x\| = r$, $\|a\| < \tilde{r}$. Then F has a zero in $\{x \in E^p \mid \|x\| \leq r\}$, and for almost all $a \in E^p$, $\|a\| < \tilde{r}$, there is a zero curve γ of*

$$\rho_a(\lambda, x) = \lambda F(x) + (1 - \lambda)(x - a),$$

along which the Jacobian matrix $D\rho_a(\lambda, x)$ has full rank, emanating from $(0, a)$ and reaching a zero $\bar{x}$ of F at $\lambda = 1$. Furthermore, γ has finite arc length if $DF(\bar{x})$ is nonsingular.

These theoretical algorithms have been implemented in production quality mathematical software packages such as PITCON [11], CONKUB [7], and HOMPACK [18]. The latter is an extensive collection of FORTRAN 77 routines implementing three different tracking algorithms for problems with both dense and sparse Jacobian matrices, and containing high level drivers for special classes of problems.

4. Basic optimization homotopies. Consider first the unconstrained optimization problem

$$\min_x f(x). \tag{3}$$

THEOREM [16]. *Let $f : E^n \to E$ be a C^3 convex map with a minimum at $\tilde{x}$, $\|\tilde{x}\|_2 \leq M$. Then for almost all a, $\|a\|_2 < M$, there exists a zero curve γ of the homotopy map*

$$\rho_a(\lambda, x) = \lambda \nabla f(x) + (1 - \lambda)(x - a),$$

along which the Jacobian matrix $D\rho_a(\lambda, x)$ has full rank, emanating from $(0, a)$ and reaching a point $(1, \tilde{x})$, where $\tilde{x}$ solves (3).

A function is called uniformly convex if it is convex and its Hessian's smallest eigenvalue is bounded away from zero. Consider next the constrained optimization problem

$$\min_{x \geq 0} f(x). \tag{4}$$

This is more general than it might appear because the general convex quadratic program reduces to a problem of the form (4).

THEOREM [16]. *Let $f : E^n \to E$ be a C^3 uniformly convex map. Then there exists $\delta > 0$ such that for almost all $a \geq 0$ with $\|a\|_2 < \delta$ there exists a zero curve γ of the homotopy map*

$$\rho_a(\lambda, x) = \lambda K(x) + (1 - \lambda)(x - a),$$

where

$$K_i(x) = -\left|\frac{\partial f(x)}{\partial x_i} - x_i\right|^3 + \left(\frac{\partial f(x)}{\partial x_i}\right)^3 + x_i^3,$$

along which the Jacobian matrix $D\rho_a(\lambda, x)$ has full rank, connecting $(0, a)$ to a point $(1, \bar{x})$, where $\bar{x}$ solves the constrained optimization problem (4).

Given $F : E^n \to E^n$, the nonlinear complementarity problem is to find a vector $x \in E^n$ such that

$$x \geq 0, \quad F(x) \geq 0, \quad x^t F(x) = 0. \tag{5}$$

At a solution $\bar{x}$, $\bar{x}$ and $F(\bar{x})$ are "complementary" in the sense that if $\bar{x}_i > 0$, then $F_i(\bar{x}) = 0$, and if $F_i(\bar{x}) > 0$, then $\bar{x}_i = 0$. This problem is difficult because there are linear constraints $x \geq 0$, nonlinear constraints $F(x) \geq 0$, and a combinatorial aspect from the complementarity condition $x^t F(x) = 0$. It is interesting that homotopy methods can be adapted to deal with nonlinear constraints and combinatorial conditions.

Define $G : E^n \to E^n$ by

$$G_i(z) = -\left|F_i(z) - z_i\right|^3 + \left(F_i(z)\right)^3 + z_i^3, \quad i = 1, \ldots, n,$$

and let

$$\rho_a(\lambda, z) = \lambda G(z) + (1 - \lambda)(z - a).$$

THEOREM [15]. *Let $F : E^n \to E^n$ be a C^2 map, and let the Jacobian matrix $DG(z)$ be nonsingular at every zero of $G(z)$. Suppose there exists $r > 0$ such that $z > 0$ and $z_k = \|z\|_\infty \geq r$ imply $F_k(z) > 0$. Then for almost all $a > 0$ there exists a zero curve γ of $\rho_a(\lambda, z)$, along which the Jacobian matrix $D\rho_a(\lambda, z)$ has full rank, having finite arc length and connecting $(0, a)$ to $(1, \bar{z})$, where $\bar{z}$ solves (5).*

THEOREM [15]. *Let $F : E^n \to E^n$ be a C^2 map, and let the Jacobian matrix $DG(z)$ be nonsingular at every zero of $G(z)$. Suppose there exists $r > 0$ such that $z \geq 0$ and $\|z\|_\infty \geq r$ imply $z_k F_k(z) > 0$*

for some index k. Then there exists $\delta > 0$ such that for almost all $a \geq 0$ with $\|a\|_\infty < \delta$ there exists a zero curve γ of $\rho_a(\lambda, z)$, along which the Jacobian matrix $D\rho_a(\lambda, z)$ has full rank, having finite arc length and connecting $(0, a)$ to $(1, \bar{z})$, where $\bar{z}$ solves (5).

Homotopy algorithms for convex unconstrained optimization are only of theoretical interest, and are generally not computationally competitive with other approaches, but it is reassuring that the globally convergent homotopy techniques can theoretically be directly applied. For constrained optimization the homotopy approach offers some advantages, and, especially for the nonlinear complementarity problem, is competitive with other algorithms. See [19] for an application of homotopy techniques to the linear complementarity problem. Constrained optimization is addressed in the next few sections.

5. Expanded Lagrangian Homotopy. The expanded Lagrangian homotopy method of Poore [9, 10] is applicable to the general nonlinear programming problem

$$\begin{aligned} &\min \theta(x) \\ \text{subject to} \quad &g(x) \leq 0, \\ &h(x) = 0, \end{aligned}$$

where $x \in E^n$, θ is real valued, g is an m-dimensional vector, and h is a p-dimensional vector. Assume that θ, g, and h are C^2. In this general situation the complete formulation and solution algorithm for the expanded Lagrangian homotopy are rather complicated. The essence of the method is presented here, referring the reader to [9] and [10] for a discussion of the theoretical and practical subtleties. The technique has been applied to linear programming [9] and the linear complementarity problem [19], but is currently primarily of theoretical interest.

The expanded Lagrangian approach may be described as an optimization/continuation approach and has in its simplest form two main steps.

Step 1. (Optimization phase).

At $r = r_0 > 0$ solve the unconstrained minimization problem

$$\min_x P(x, r)$$

where

$$P(x, r) = \theta(x) + \frac{1}{2r} h(x)^t h(x) - r \sum_{i=1}^{m} \ln\bigl(-g_i(x)\bigr).$$

Step 2A. (Switch to expanded system).

A (local) solution of $\min P$ must satisfy

$$0 = \nabla_x P = \nabla\theta(x) + \frac{h(x)^t \nabla h(x)}{r} - \sum_{i=1}^{m} \frac{r}{g_i(x)} \nabla g_i(x).$$

Introduce the following variables:

$$\begin{aligned} \beta &= \frac{h(x)}{r}, \\ \mu_i &= \frac{r}{-g_i(x)}, \qquad i = 1, \ldots, m, \end{aligned}$$

which ultimately represent the Lagrange multipliers. This helps to remove the inevitable ill-conditioning associated with penalty methods for small r and we thus obtain our equivalent but expanded system:

$$\begin{aligned} \nabla\theta(x) + \beta^t \nabla h(x) + \mu^t \nabla g(x) &= 0, \\ h(x) - r\beta &= 0, \\ \mu_i g_i(x) + r &= 0, \qquad i = 1, \ldots, m. \end{aligned}$$

(Remark. As a result of the optimization phase and the initial starting point with $r_0 > 0$, the solution $x^{(0)}$ of $\min P(x, r_0)$ satisfies $g(x^{(0)}) < 0$. As a consequence, $\mu^{(0)} > 0$ from the definition of μ. μ remains positive until $r = 0$ where we formally have

$$\begin{aligned} \nabla\theta(x) + \beta^t \nabla h(x) + \mu^t \nabla g(x) &= 0, \\ h(x) &= 0, \\ g(x) &\leq 0, \\ \mu &\geq 0, \\ \mu_i g_i(x) &= 0, \qquad i = 1, \ldots, m, \end{aligned}$$

which implies that we have solved the problem.)

In practice we do not solve the optimization problem $\min P$ to high accuracy since a highly accurate solution may have only a digit or two in common with the final answer. However, it is imperative that ∇P be reasonably small in magnitude, say less than $r_0/10$. The expanded system is converted to a homotopy map by letting $r = r_0(1-\lambda)$ and modifying the first equation to obtain:

$$\begin{aligned} \nabla\theta(x) + \beta^t \nabla h(x) + \mu^t \nabla g(x) - \frac{r}{r_0} \nabla P(x^{(0)}, r_0) &= 0, \\ h(x) - r\beta &= 0, \\ \mu_i g_i(x) + r &= 0, \qquad i = 1, \ldots, m. \end{aligned} \tag{6}$$

Write this system of $n + p + m$ equations in the $n + p + m + 1$ variables λ, x, β, μ as

$$\Upsilon(\lambda, x, \beta, \mu) = 0.$$

Step 2B. (Track the zero curve of Υ from $r = r_0$ to $r = 0$.)

Starting with arbitrary $r_0 > 0$ and feasible interior point $x^{(0)}$ $(g(x^{(0)}) < 0)$, the rest of the initial point $(0, x^{(0)}, \beta^{(0)}, \mu^{(0)})$ is given by

$$\begin{aligned} \beta^{(0)} &= \frac{h(x^{(0)})}{r_0}, \\ \mu_i^{(0)} &= \frac{r_0}{-g_i(x^{(0)})}, \qquad i = 1, \ldots, m. \end{aligned}$$

This approach requires careful attention to implementation details. For example, the linear algebra and globalization techniques with dynamic scaling are critically important in the optimization phase. For degenerate problems the path can still be long. One possible resolution is the use of shifts and weights as developed in the method of multipliers [3], but holding $r = r_0$ fixed. (This approach is currently under investigation in the context of linear programming [9].) Note that the

optimization phase (Step 1) can be omitted altogether, starting Step 2B with an arbitrary interior feasible point $x^{(0)}$ ($g(x^{(0)}) < 0$), so that (6) is a true global homotopy. As a practical matter, however, it is advantageous to get a good starting point by doing Step 1 with a small r_0.

6. An example of a special purpose homotopy. The optimization problem that we consider here is to maximize the lowest eigenvalue η of the eigenvalue problem

$$A(v)u - \eta\, B(v)u = 0, \tag{7}$$

where A and B are symmetric positive definite matrices which depend on a vector of design variables v with components v_i. Equation (7) can represent various engineering problems such as buckling and vibration problems. In a typical problem the design variables have lower and upper bounds and have some cost associated with them. Since the lowest eigenvalue of (7) may be written as the minimum of the Rayleigh quotient

$$\eta_1 = \min_u \frac{u^T A u}{u^T B u}, \tag{7a}$$

the optimization problem may be formulated as

$$\begin{aligned} &\max_v \min_u \frac{u^T A u}{u^T B u} \\ &\text{such that} \quad c^T v - \theta = 0 \\ &\qquad \text{and} \quad v_{i\,min} \le v_i \le v_{i\,max} \quad \text{for } i = 1, \ldots, M, \end{aligned} \tag{7b}$$

where c is a positive cost vector, and θ is the amount of available resources.

A typical optimization method, applied to solve this problem, starts from a given design and continuously searches for better designs until it finds an optimum design. The trial designs along the path are of no value. Reference [13] proposed instead a method which proceeds along a path of optimal designs for increasing amounts of resource θ. The resource θ is varied between the minimum θ_{min} required to satisfy the lower bound constraints and a maximum θ_{max} when all variables are at their upper bounds.

The path consists of several smooth segments, each segment being characterized by a set I_A of variables which are at their upper or lower bounds. Along each segment, some inequality constraints can be treated as equality constraints,

$$v_j = v_{j\,min} \quad \text{or} \quad v_j = v_{j\,max} \quad \text{for } j \in I_A, \tag{8}$$

so that these variables can be eliminated from the optimization problem, while the other variables do not have to be constrained. The optimization problem along a segment can, therefore, be written as

$$\begin{aligned} &\max_{v_i} \min_u \frac{u^T A u}{u^T B u} \quad \text{for } i \notin I_A \\ &\text{such that} \quad c^T v - \theta = 0. \end{aligned} \tag{9}$$

The solution of the problem consists of three related problems: solving the optimization problem along a segment, locating the end of the segment where the set I_A changes, and finding the set I_A for the next segment.

It is common practice to normalize the eigenvector u such that the denominator of Rayleigh's quotient is unity and to treat this as an equality constraint. Then, using Lagrange multipliers η and μ, the augmented function P^* is formed:

$$P^* = u^T A u - \eta \left[u^T B u - 1\right] - \mu \left[c^T v - \theta\right]. \tag{10}$$

The following stationary conditions are obtained by taking the first derivative of P^* with respect to v_i, u, η, and μ, and setting it equal to zero:

i) Optimality conditions

$$u^T \frac{\partial A}{\partial v_i} u - \eta u^T \frac{\partial B}{\partial v_i} u - \mu c_i = 0 \quad \text{for } i \notin I_A. \tag{11}$$

ii) Eigenvalue problem

$$Au - \eta B u = 0. \tag{12}$$

iii) Normalization constraint

$$1 - u^T B u = 0. \tag{13}$$

iv) Total resource constraint

$$\theta - c^T v = 0. \tag{14}$$

Equations (11)-(14) form a system of nonlinear equations to be solved for v_i, u, η, and μ. A homotopy method is used to find the solutions of these equations as a function of θ.

In certain ranges of structural resources, the optimal solution is known to be bimodal, i.e., the lowest eigenvalue is repeated. The formulation for bimodal solutions is given in the appendix of [13]. The existence of bimodal solutions also introduces additional transitions (bimodal to unimodal and vice versa) along the path of optimum solutions.

The homotopy method as described here earlier is intended to solve a *single* nonlinear system of equations, and converge from an arbitrary starting point with probability one. In this context $\theta \in [0, 1]$, and the zero curve γ is bounded and leads to the (single) desired solution at $\theta = 1$. The a vector, viewed as an artificial perturbation of the problem, plays a crucial role. In the version of the method employed here, $\theta \in (\theta_0, \theta_1)$, *each* point along γ has physical significance, and a is fixed at zero (no perturbation). Because a is not random, the claimed properties for γ hold only in subintervals (θ_0, θ_1) of $[0, \infty)$. Detecting and dealing with these subinterval transition points is the essence of the modification of the homotopy method used in this section.

Switching from one segment to the next

There are four types of events which end a segment and start a new one:
Type 1: a bound constraint becoming active (i.e., being satisfied as an equality);
Type 2: a bound constraint becoming inactive;
Type 3: transition from a unimodal solution to a bimodal solution;
Type 4: transition from a bimodal solution to a unimodal solution.

To switch from one segment to the next, we first need to locate the transition point. At a transition point there are a number of solution paths which satisfy the stationary equations, and we need to choose the optimum path.

Transition points are located by checking the bound constraints and the optimality conditions. The bound constraints

$$v_{i\,min} \leq v_i \leq v_{i\,max} \quad \text{for } i = 1, \ldots, M \tag{15}$$

are checked to detect a transition point of type 1.

Optimality of the solution is checked by the Kuhn-Tucker conditions and the second-order conditions discussed below. The solution satisfies the Kuhn-Tucker conditions when all Lagrange multipliers are nonnegative. So a transition of type 2 is detected by checking the positivity of the Lagrange multipliers associated with the bound constraints. These multipliers are obtained by adding the bound constraints to the formulation (9) and replacing the augmented function P^* by

$$P^* = u^T A u - \eta \left[u^T B u - 1\right] - \mu \left[c^T v - \theta\right] - \sum_{i \in I_A} \lambda_{1i} \left[v_{i\,min} - v_i\right] - \sum_{i \in I_A} \lambda_{2i} \left[v_i - v_{i\,max}\right]. \quad (16)$$

Taking the first derivative of P^* with respect to v_i gives

$$u^T \frac{\partial A}{\partial v_i} u - \eta u^T \frac{\partial B}{\partial v_i} u - \mu c_i + \lambda_{1i} - \lambda_{2i} = 0 \quad \text{for } i \in I_A. \quad (17)$$

Since λ_{1i} is 0 for $v_i \neq v_{i\,min}$ and λ_{2i} is 0 for $v_i \neq v_{i\,max}$ for the above equations, λ_{1i} and λ_{2i} are given by

$$\begin{aligned} \lambda_{1i} &= -u^T \frac{\partial A}{\partial v_i} u + \eta u^T \frac{\partial B}{\partial v_i} u + \mu c_i \quad \text{for } v_i = v_{i\,min}, \\ \lambda_{2i} &= u^T \frac{\partial A}{\partial v_i} u - \eta u^T \frac{\partial B}{\partial v_i} u - \mu c_i \quad \text{for } v_i = v_{i\,max}. \end{aligned} \quad (18)$$

A type 2 transition is detected by a Lagrange multiplier becoming nonpositive. Similar equations for the bimodal case are given in the appendix of [13].

The bimodal formulation replaces η by η_1 and η_2 which are the Lagrange multipliers for the normalization constraints on the two repeated eigenvectors. When one of them becomes negative, the corresponding eigenvector should be removed for the optimum design, so that we have a transition of type 4 from bimodal to unimodal design.

For a transition of type 3, we need to check if there is another eigenvector associated with a lower eigenvalue. This can be accomplished by checking the second-order optimality conditions for the eigenvector variables u given by

$$r^T \left[\nabla_u^2 P^*\right] r > 0 \quad \text{for every } r \text{ such that } \quad (\nabla_u h)^T r = 0, \quad (19)$$

where

$$\begin{aligned} \left[\nabla_u^2 P^*\right] &= \left[\frac{\partial^2 P^*}{\partial u_s \partial u_t}\right], \\ \nabla_u h &= \left[\frac{\partial h}{\partial u_s}\right], \\ h &= u^T B u - 1. \end{aligned}$$

Alternatively we can solve the eigenvalue problem (12) for the current design and check whether the buckling load obtained from the stationary conditions is truly the lowest one. The transition of type 3 is detected by checking if

$$\eta \neq \eta_1, \quad (20)$$

where η is the solution obtained from the stationary conditions while η_1 is the lowest eigenvalue obtained by solving the eigenvalue problem (12) for the set of design variables v obtained from the stationary conditions.

Once a transition point is located, we need to choose a path which satisfies the optimality conditions. Choosing an optimum path constitutes finding a set of active bound constraints for type 1 and 2 transitions and the correct eigenvectors for type 3 and 4 transitions. These are obtained by using the Lagrange multipliers of the previous path and the sensitivity calculation on the buckling load. The procedure is explained separately for each type of transition.

A type 1 transition occurs when one of design variables, v_i, hits the upper or lower bound. Then v_i is set at $v_{i\,max}$ or $v_{i\,min}$ and treated as a constant value. The number of design variables is reduced by one.

At a type 2 transition, one of the Lagrange multipliers for the bound constraints, λ_{1i} and λ_{2i}, is found to be negative. The bound constraint corresponding to the negative λ_{1i} or λ_{2i} is set to be inactive and the number of design variables is increased by one.

At a transition from a unimodal solution to a bimodal solution (a type 3 transition), the formulation requires two eigenvectors, u_1 and u_2, for the solution of the upcoming bimodal path. These eigenvectors can be obtained by solving the eigenproblem (12) of the previous unimodal formulation, since the lowest eigenvalue is repeated at the bimodal transition point.

At a transition from a bimodal to a unimodal solution (a type 4 transition), two eigenvectors are given from the bimodal solution. One of the Lagrange multipliers for the normalization constraints, η, is known to be negative from the previous transition check, so the eigenvector corresponding to the positive η is chosen.

Some of the above transitions can occur simultaneously. Special treatment is required in certain cases where the Lagrange multipliers are not available. In general, the optimum design requires at least one design variable v_i for a unimodal case and two design variables for a bimodal case. At a type 1 transition, the number of design variables is reduced by one, and at a type 3 transition the bimodal formulation requires one more design variable in case the previous unimodal path has only one design variable. So some type 1 or type 3 transitions occur simultaneously with a type 2 transition which allows an additional design variable. In that case, the Lagrange multipliers λ_{1i} and λ_{2i}, which are used at a type 2 transition to determine a new design variable, are not available. We then rely on the sensitivity information of η with respect to v. For a unimodal case, the location of the new design variable v_i is determined where $d\eta/d\theta$ is maximized. For a bimodal case, we need to find a combination of i and j which maximizes the value of the eigenvalue for a small increment of the total available resource. Considering the bound constraints in the formulation, the new design variables are determined by

$$\max_{i,j} \quad \frac{d\eta}{d\theta} = \frac{\partial \eta_1}{\partial v_i}\frac{dv_i}{d\theta} + \frac{\partial \eta_1}{\partial v_j}\frac{dv_j}{d\theta} \tag{21}$$

$$\begin{aligned}
\text{such that } & \frac{\partial \eta_1}{\partial v_i}\frac{dv_i}{d\theta} + \frac{\partial \eta_1}{\partial v_j}\frac{dv_j}{d\theta} = \frac{\partial \eta_2}{\partial v_i}\frac{dv_i}{d\theta} + \frac{\partial \eta_2}{\partial v_j}\frac{dv_j}{d\theta} \\
& \frac{dv_i}{d\theta} \geq 0 \quad \text{for } v_i = v_{i\,min} \\
& \frac{dv_i}{d\theta} \leq 0 \quad \text{for } v_i = v_{i\,max} \\
& \frac{dv_j}{d\theta} \geq 0 \quad \text{for } v_j = v_{j\,min} \\
\text{and } & \frac{dv_j}{d\theta} \leq 0 \quad \text{for } v_j = v_{j\,max}
\end{aligned}$$

where η_1 and η_2 are the eigenvalues corresponding to the eigenvectors u_1 and u_2, respectively.

After we obtain the design variables v and the eigenvectors u, we need the Lagrange multipliers μ and η at the transition point to complete the set of starting values for the next solution path. These are obtained by solving the stationary conditions for the given u and v. For example, in the unimodal case, η is obtained from the stability conditions (12) and μ is obtained by solving one of the optimality conditions (11).

In [13] this procedure was applied to the design of a foundation which supports a column for maximum buckling load, where the total available foundation was used as a homotopy parameter. Starting from a minimum foundation which satisfies the lower bound (in this case zero), a set of optimum foundation designs was obtained for the full range of total foundation stiffness. Special purpose homotopies such as the one described in this section can be very efficient, and are definitely competitive with other algorithms.

7. Example of a smooth envelope function for nonlinear constraints. The two previous sections presented ways to deal with inequality constraints. Both are theoretically "correct" and computationally "practical". However, there are numerous practical difficulties in dealing with them, and the implementation and tuning details become absolutely crucial. e.g., with the expanded Lagrangian formulation, line searches may generate negative arguments for the ln functions, and the homotopy zero curve may diverge if the Step 1 solution is not good enough. For the approach in Section 6, the detection and switching criteria for transition points may become extremely cumbersome and inefficient. This section suggests an alternate way of dealing with inequality constraints.

Consider inequality constraints of the form

$$g_i(x) \le 0, \quad i = 1, \ldots, m, \tag{22}$$

where each $g_i : E^n \to E$ is C^2. For a constant $\rho > 0$, the Kreisselmeier-Steinhauser [5] envelope function for (22) is

$$K(x) = \frac{1}{\rho} \ln \left[\sum_{i=1}^{m} \exp\big(\rho g_i(x)\big) \right]. \tag{23}$$

$K(x)$ is a cumulative measure of the satisfaction or violation of the constraints (22). Let $g_{max}(x) = \max\{g_1(x), \ldots, g_m(x)\}$, and observe that

$$K(x) = g_{max}(x) + \frac{1}{\rho} \ln \left[\sum_{i=1}^{m} \exp\Big(\rho\big(g_i(x) - g_{max}(x)\big)\Big) \right], \tag{24}$$

from which it directly follows [2] that

$$g_{max}(x) \le K(x) \le g_{max}(x) + \frac{1}{\rho} \ln m. \tag{25}$$

Thus the envelope $K(x)$ follows the maximum constraint, more closely for large ρ. In particular, (22) could be replaced by

$$K(x) \le 0 \tag{26}$$

with an error of no more than $(\ln m)/\rho$.

The choice of ρ involves a tradeoff between modelling the maximum constraint (large ρ preferred) and avoiding large gradients (small ρ preferred). If the practical criterion for an active constraint is $|g_i| \le \epsilon$, then a choice for ρ which has worked well in practice [2] is

$$\rho = \frac{\ln m}{\epsilon}. \tag{27}$$

Observe that $K(x)$ is C^2 and defined *everywhere*, a decided advantage over barrier functions. Furthermore, (26) is a *single* nonlinear constraint, which makes any active set strategy very simple. (26) has been successfully used in large scale structural optimization [2] and optimal control [5].

8. Probability-one homotopy for Kuhn-Tucker optimality conditions. Section 7 explained why the approaches of Sections 5 and 6 are not always entirely adequate. The cumulative constraint function (23) is however decidedly unnatural, extremely nonlinear and ill conditioned for large ρ, and does not take advantage of a known solution to a related problem. Consider again the general nonlinear programming problem of Section 4:

$$\begin{aligned} &\min \theta(x) \\ \text{subject to} \quad &g(x) \leq 0, \\ &h(x) = 0, \end{aligned} \tag{28}$$

under the same assumptions mentioned before. The Kuhn-Tucker necessary optimality conditions for (28) are

$$\begin{aligned} \nabla\theta(x) + \beta^t \nabla h(x) + \mu^t \nabla g(x) &= 0, \\ h(x) &= 0, \\ g(x) &\leq 0, \\ \mu &\geq 0, \\ \mu^t g(x) &= 0, \end{aligned} \tag{29}$$

where $\beta \in E^p$ and $\mu \in E^m$. Following Mangasarian [6] and Watson [15], the complementarity conditions $\mu \geq 0$, $g(x) \leq 0$, $\mu^t g(x) = 0$ are replaced by the equivalent nonlinear system of equations

$$W(x, \mu) = 0, \tag{30a}$$

where

$$W_i(x, \mu) = -\left|\mu_i + g_i(x)\right|^3 + \mu_i^3 - \left(g_i(x)\right)^3, \quad i = 1, \ldots, m. \tag{30b}$$

Thus the optimality conditions (29) take the form

$$F(x, \beta, \mu) = \begin{pmatrix} \left[\nabla\theta(x) + \beta^t \nabla h(x) + \mu^t \nabla g(x)\right]^t \\ h(x) \\ W(x, \mu) \end{pmatrix} = 0. \tag{31}$$

With $z = (x, \beta, \mu)$, the proposed homotopy map is

$$\rho_a(\lambda, z) = \lambda F(z) + (1 - \lambda)(z - a), \tag{32}$$

where $a \in E^{n+p+m}$. Simple conditions on θ, g, and h guaranteeing that the above homotopy map $\rho_a(\lambda, z)$ will work are unknown, although this map has worked very well on some difficult fuel optimal orbital rendezvous problems [23].

Frequently in practice the functions θ, g, and h involve a parameter vector c, and a solution to (28) is known for some $c = c^{(0)}$. Suppose that the problem under consideration has parameter vector $c = c^{(1)}$. Then

$$c = (1 - \lambda)c^{(0)} + \lambda c^{(1)} \tag{33}$$

parametrizes c by λ and $\theta = \theta(x; c) = \theta(x; c(\lambda))$, $g = g(x; c(\lambda))$, $h = h(x; c(\lambda))$. The optimality conditions in (31) become functions of λ as well, $F(\lambda, x, \beta, \mu) = 0$, and

$$\rho_a(\lambda, z) = \lambda F(\lambda, z) + (1 - \lambda)(z - a) \tag{34}$$

is a highly implicit nonlinear function of λ. If $F(0, z^{(0)}) = 0$, a good choice for a in practice has been found to be $a = z^{(0)}$. A natural choice for a homotopy would be simply

$$F(\lambda, z) = 0, \tag{35}$$

since the solution $z^{(0)}$ to $F(0, z) = 0$ (the problem corresponding to $c = c^{(0)}$) is known. However, for various technical reasons, (34) is much better than (35) [23].

The homotopy (34) was used in [23] to solve a fuel-optimal orbital rendezvous problem, and for such optimal control problems appears to be far superior to other known algorithms.

9. Acknowledgement. The work of L. T. Watson was supported in part by AFOSR Grant 85–0250, and that of R. T. Haftka by NASA grant NAG-1-168.

10. References.

[1] E. Allgower and K. Georg, *Simplicial and continuation methods for approximating fixed points*, SIAM Rev., 22 (1980), pp. 28–85.

[2] J.-F. M. Barthelemy and M. F. Riley, *Improved multi-level optimization approach for the design of complex engineering systems*, AIAA J., 26 (1988), pp. 353–360.

[3] D. P. Bertaekas, *Constrained Optimization and Lagrange Multiplier Methods*, Academic Press, New York, 1982.

[4] S. N. Chow, J. Mallet-Paret, and J. A. Yorke, *Finding zeros of maps: Homotopy methods that are constructive with probability one*, Math. Comput., 32 (1978), pp. 887–899.

[5] G. Kreisselmeier and R. Steinhauser, *Systematic control design by optimizing a vector performance index*, Proc. IFAC Symp. on Computer Aided Design of Control Systems, Zurich, Switzerland, (1979), pp. 113–117.

[6] O.L. Mangasarian, *Equivalence of the complementarity problem to a system of nonlinear equations*, SIAM Journal on Applied Mathematics, 31 (1976), pp. 89–92.

[7] R. Mejia, *CONKUB: A conversational path-follower for systems of nonlinear equations*, J. Comput. Phys., 63 (1986), pp. 67–84.

[8] A.P. Morgan, *Solving polynomial systems using continuation for scientific and engineering problems*, Prentice-Hall, Englewood Cliffs, NJ, (1987).

[9] A. B. Poore and D. Soria, *Continuation algorithms for linear programming*, in preparation.

[10] A. B. Poore and Q. Al-Hassan, *The expanded Lagrangian system for constrained optimization problems*, SIAM J. Control Optim., to appear.

[11] W. C. Rheinboldt and J. V. Burkardt, *Algorithm 596: A program for a locally parameterized continuation process*, ACM Trans. Math. Software, 9 (1983), pp. 236-241.

[12] H. Scarf, *The Computation of Economic Equilibria*, Yale University Press, (1973).

[13] Y. S. Shin, R. T. Haftka, L. T. Watson, and R. H. Plaut, *Tracing structural optima as a function of available resources by a homotopy method*, Comput. Methods Appl. Mech. Engrg., 70 (1988), pp. 151–164.

[14] L. T. Watson, *A globally convergent algorithm for computing fixed points of C^2 maps*, Appl. Math. Comput., 5 (1979), pp. 297–311.

[15] L.T. Watson, *Solving the nonlinear complementarity problem by a homotopy method*, SIAM J. Control Optim., 17,1 (1979), pp. 36–46.

[16] L. T. WATSON, *Computational experience with the Chow-Yorke algorithm*, Math. Programming, 19 (1980), pp. 92–101.

[17] L. T. WATSON, *Numerical linear algebra aspects of globally convergent homotopy methods*, Tech. Report TR–85–14, Dept. of Computer Sci., VPI&SU, Blacksburg, VA, 1985, and SIAM Rev., 28 (1986), pp. 529–545.

[18] L. T. WATSON, S. C. BILLUPS, AND A. P. MORGAN, *HOMPACK: A suite of codes for globally convergent homotopy algorithms*, Tech. Rep. 85–34, Dept. of Industrial and Operations Eng., Univ. of Michigan, Ann Arbor, MI, 1985, and ACM Trans. Math. Software, 13 (1987), pp. 281–310.

[19] L. T. WATSON, J. P. BIXLER, AND A. B. POORE, *Continuous homotopies for the linear complementarity problem*, Tech. Report TR–87–38, Dept. of Computer Sci., VPI&SU, Blacksburg, VA, 1987.

[20] L. T. WATSON AND D. FENNER, *Chow-Yorke algorithm for fixed points or zeros of C^2 maps*, ACM Trans. Math. Software, 6 (1980), pp. 252–260.

[21] L. T. WATSON AND C. Y. WANG, *A homotopy method applied to elastica problems*, Internat. J. Solids Structures, 17 (1981), pp. 29–37.

[22] L. T. WATSON AND W. H. YANG, *Optimal design by a homotopy method*, Applicable Anal., 10 (1980), pp. 275–284.

[23] G. VASUDEVAN, L. T. WATSON, AND F. H. LUTZE, *A homotopy approach for solving constrained optimization problems*, Tech. Rep, Dept. of Computer Sci., VPI&SU, Blacksburg, VA, 1988.