H

Hamilton–Jacobi–Bellman Equation

WILLIAM R. ESPOSITO
Department Chemical Engineering,
Princeton University, Princeton, USA

MSC2000: 49L20, 34H05, 90C39

Article Outline

Keywords

Dynamic programming; Continuous-time optimal control; Hamilton–Jacobi–Bellman equation

Even though *dynamic programming* [2] was originally developed for the solution of problems which exhibit discrete types of decisions, it has also been applied to continuous formulations. In this article, the application of dynamic programming to the solution of continuous-time optimal control problems is discussed. By discretizing the problem, applying the dynamic programming equations, then returning to the continuous domain, a partial differential equation results, the *Hamilton–Jacobi–Bellman equation* (HJB equation). This equation is often referred to as the *continuous-time equivalent of the dynamic programming algorithm*. In this article, the HJB equation will first be derived. A simple application will be presented, in addition to its use in solving the linear quadratic control problem. Finally, a brief overview of some solution methods and applications presented in the literature will be given.

Problem Formulation

The dynamic programming approach will be applied to a system of the following form:

$$\begin{cases} \dot{z}(t) = f(z(t), u(t)), \\ z(0) = z_0, \qquad\qquad 0 \le t \le T, \end{cases} \tag{1}$$

where $z(\mathrm{t}) \in \mathbf{R}^n$ is the state vector at time t with time derivative given by $\dot{z}(t)$, $u(t) \in U \subset \mathbf{R}^m$ is the control vector at time t, U is the set of control constraints, and T is the terminal time. The function $f(z(t), u(t))$ is continuously differentiable with respect to z and continuous with respect to u. The set of admissible control trajectories are given by the piecewise constant functions, $\{u(t)$: $u(t) \in U, \forall t \in [0, T]\}$. It is assumed that for any admissible control trajectory, that a state trajectory $z^u(t)$ exists and is unique.

The objective is to determine a control trajectory and the corresponding state trajectory which minimizes a cost function of the form:

$$h(z^u(T)) + \int_0^T g(z^u(t), u(t))\, dt, \tag{2}$$

where the functions g, and h are continuously differentiable with respect to both z and u.

Derivation

The derivation of the Hamilton–Jacobi–Bellman equation is taken from [3]. The time horizon is first dis-

cretized into N equally spaced intervals with:

$$\delta = \frac{T}{N}.$$

Also, the state and control are represented by:

$$z_k = z(k\delta), \quad k = 0, \ldots, N,$$
$$u_k = u(k\delta), \quad k = 0, \ldots, N.$$

The continuous-time system is approximated by:

$$z_{k+1} = z_k + f(z_k, u_k)\delta.$$

The cost function is rewritten as:

$$h(z_N) + \sum_{k=0}^{N-1} g(z_k, u_k)\delta.$$

The dynamic programming algorithm is now applied with the following definitions:

- $J^*(t, z)$ is the optimal cost-to-go for the continuous problem;
- $\widehat{J}^*(t, z)$ is the optimal cost-to-go for the discrete approximation.

The dynamic programming equations then take the form:

$$\widehat{J}^*(N\delta, z) = h(z), \tag{3}$$

$$\widehat{J}^*(k\delta, z) = \min_{u\in U}\left[g(z, u)\delta + \widehat{J}^*((k+1)\delta, z + f(z, u)\delta)\right], \quad k = 0, \ldots, N-1. \tag{4}$$

It is assumed that $\widehat{J}^*(t, z)$ has the necessary differentiability requirements to write the following Taylor series expansion:

$$\widehat{J}^*((k+1)\delta, z + f(z, u)\delta) = \widehat{J}^*(k\delta, z) + \nabla_t \widehat{J}^*(k\delta, z)\delta + \nabla_z \widehat{J}^{*\top}(k\delta, z) f(z, u)\delta + o(\delta), \tag{5}$$

where $o(\delta)$ represents second order terms which satisfy $o(\delta)/\delta \to 0$ as $\delta \to 0$. Substituting (5) into (4) results in:

$$\widehat{J}^*(k\delta, z) = \min_{u\in U}\left[g(z, u)\delta + \widehat{J}^*(k\delta, z) + \nabla_t \widehat{J}^*(k\delta, z)\delta + \nabla_z \widehat{J}^{*\top}(k\delta, z) f(z, u)\delta + o(\delta)\right]. \tag{6}$$

Dividing (6) by δ and $\widehat{J}^*(k\delta, z)$, and taking the limit as $\delta \to 0$ with the assumption that

$$\lim_{\substack{k\to\infty \\ \delta\to 0 \\ k\delta = t}} \widehat{J}^*(k\delta, z) = J^*(t, z)$$

results in

$$0 = \min_{u\in U}\left[g(z, u) + \nabla_t J^*(t, z) + \nabla_x J^{*\top}(t, z) f(z, u)\right], \quad \forall t, z, \tag{7}$$

with the boundary condition

$$J^*(T, z) = h(z).$$

This partial differential equation is known as the Hamilton–Jacobi–Bellman equation (HJB equation).

Sufficiency Theorem

This theorem is presented in [3]. Suppose $V(t, z)$ is a solution to the HJB equation, that is, V is continuously differentiable with respect to z and t and satisfies:

$$0 = \min_{u\in U}\left[g(z, u) + \nabla_t V(t, z) + \nabla_x V^{\top}(t, z) f(z, u)\right], \quad \forall z, t, \tag{8}$$

$$V(T, z) = h(z), \quad \forall z. \tag{9}$$

Suppose also that $\mu^*(t, z)$ attains the minimum in (8) for all t and z. Let $z^*(t)$ be the state trajectory obtained from the given initial condition $z(0)$ when the control trajectory $u^*(t) = \mu^*(t, z^*(t))$ is used. (That is, $z^*(0) = z(0), \dot{z}^* = f(z^*(t), \mu^*(t, z^*(t)))$; one also assumes that this differential equation has a unique solution starting at any pair (t, z) and that the control trajectory is piecewise continuous in time.) Then V is the unique solution of the HJB equation and is equal to the optimal cost-to-go function

$$V(t, z) = J^*(t, z), \quad \forall z, t.$$

Furthermore, the control trajectory, $u^*(t)$ is optimal for all $t \in [0, T]$.

Example

Consider the simple dynamic system:

$$\dot{z}(t) = u(t)$$

with the control bounded by $u(t) \in [-1, 1]$ and time over the range $t \in [0, T]$. The cost function is given as:

$$\frac{1}{2} z(T)^2.$$

Writing the HJB equation for this system gives

$$0 = \min_{u \in [-1,1]} \left[\nabla_t V(t,z) + \nabla_z V(t,z) u \right], \quad \forall t, z,$$

with the boundary condition,

$$V(T,z) = \frac{1}{2} z^2.$$

The obvious choice of a control policy is to drive the state to zero as fast as possible and keep it there. This corresponds to the policy:

$$\mu^*(t,z) = -\operatorname{sgn}(z) = \begin{cases} 1 & \text{if } z < 0, \\ 0 & \text{if } z = 0, \\ -1 & \text{if } z > 0. \end{cases}$$

The cost associated with this policy for a given initial time and state is:

$$J^*(t,z) = \frac{1}{2} \left(\max\{0, |z| - (T-t)\} \right)^2.$$

This function satisfies the terminal condition $J^*(T,z) = z^2/2$. Also,

$$\nabla_t J^*(t,z) = \max\{0, |z| - (T-t)\},$$
$$\nabla_z J^*(t,z) = \operatorname{sgn}(z) \max\{0, |z| - (T-t)\}.$$

Substituting these expressions into the HJB equation results in

$$0 = \min_{u \in [-1,1]} \left[1 + \operatorname{sgn}(z) u \right] \max\{0, |z| - (T-t)\},$$

which can be shown to hold for all (t,z). The minimum is attained for $u = -\operatorname{sgn}(z)$, and one therefore concludes from the sufficiency theorem presented above that $J^*(t,z)$ is indeed the optimal cost-to-go function.

Linear-Quadratic Problem

Consider a general n-dimensional time-invariant linear system

$$\dot{z}(t) = Az(t) + Bu(t)$$

with a cost function defined by

$$z^\top(T) Q_T z(T) + \int_0^T z^\top(t) Q z(t) + u^\top(t) R u(t)\, dt,$$

where the matrices Q and Q_T are symmetric positive semidefinite, and the matrix R is symmetric positive definite. The HJB equation is written as

$$0 = \min_{u \in \mathbf{R}^m} \left[z^\top Q z + u^\top R u + \nabla_t V(t,z) + \nabla_z V^\top(t,z)(Az + Bu) \right], \quad V(T,z) = z^\top Q_T z. \tag{10}$$

Try a solution of the form:

$$V(t,z) = z^\top K(t) z,$$

where $K(t)$ is a symmetric $n \times n$ matrix. One then has

$$\nabla_z V(t,z) = 2K(t)z,$$
$$\nabla_t V(t,z) = z^\top \dot{K}(t) z.$$

Substituting the above expressions into (10) results in

$$0 = \min_{u \in \mathbf{R}^m} \left[z^\top Q z + u^\top R u + z^\top \dot{K}(t) z + 2 z^\top K(t) A z + 2 z^\top K(t) B u \right]. \tag{11}$$

The minimum is obtained when the gradient with respect to u is zero. This results in

$$2B^\top K(T) z + 2Ru = 0$$

or

$$u = -R^{-1} B^\top K(t) z.$$

Substituting this expression into (11), the following results:

$$0 = z^\top \left(\dot{K}(t) + K(t)A + A^\top K(t) - K(t) B R^{-1} B^\top K(t) + Q \right) z.$$

Therefore, $K(t)$ must satisfy the following matrix differential equation:

$$\dot{K}(t) = -K(t)A - A^\top K(t) + K(t) B R^{-1} B^\top K(t) - Q,$$

with the terminal condition

$$K(T) = Q_T.$$

This equation is known as the *continuous-time Riccati equation*.

Solution Methods and Applications

In the general case of a nonlinear system, the solution can not be determined analytically and numerical methods need to be relied on. The numerical solution of the Hamilton–Jacobi–Bellman equation is not trivial due to its partial differential nature. Additionally the HJB equation and accompanying numerical methods have been used to solve a wide variety of problems.

See [4] for many applications in the area of optimal control, and for an advocate solution by the *method of characteristics*. This classical technique for the solution of partial differential equations can be found in many textbooks. See [6] for remarks about the application of the HJB equation to minimum time optimal control problems. See [1] for an approximate method for the solution of the time-invariant HJB equation. The method consists of a reduction to a set on linear partial differential equations and an approximation via the *Galerkin spectral method*. It also presents an extensive review of various approximation approaches and an application for the voltage regulation of a power generator. See [7] for an alternating direction algorithm for the solution of HJB equations. See [5] for an application for the optimal path timing of robot manipulators and for the approximate solution of the resulting HJB equation using *finite difference methods*.

The aforementioned references are a subset of the various solution methods for and applications of the Hamilton–Jacobi–Bellman equation.

See also

- ▶ Control Vector Iteration
- ▶ Duality in Optimal Control with First Order Differential Equations
- ▶ Dynamic Programming: Average Cost Per Stage Problems
- ▶ Dynamic Programming in Clustering
- ▶ Dynamic Programming: Continuous-time Optimal Control
- ▶ Dynamic Programming: Discounted Problems
- ▶ Dynamic Programming: Infinite Horizon Problems, Overview
- ▶ Dynamic Programming: Inventory Control
- ▶ Dynamic Programming and Newton's Method in Unconstrained Optimal Control
- ▶ Dynamic Programming: Optimal Control Applications
- ▶ Dynamic Programming: Stochastic Shortest Path Problems
- ▶ Dynamic Programming: Undiscounted Problems
- ▶ High-order Maximum Principle for Abnormal Extremals
- ▶ Infinite Horizon Control and Dynamic Games
- ▶ MINLP: Applications in the Interaction of Design and Control
- ▶ Multi-objective Optimization: Interaction of Design and Control
- ▶ Multiple Objective Dynamic Programming
- ▶ Neuro-dynamic Programming
- ▶ Optimal Control of a Flexible Arm
- ▶ Optimization Strategies for Dynamic Systems
- ▶ Pontryagin Maximum Principle
- ▶ Robust Control
- ▶ Robust Control: Schur Stability of Polytopes of Polynomials
- ▶ Semi-infinite Programming and Control Problems
- ▶ Sequential Quadratic Programming: Interior Point Methods for Distributed Optimal Control Problems
- ▶ Suboptimal Control

References

1. Beard RW, Saridis GN, Wen JT (1998) Approximate solutions to the time-invariant Hamilton–Jacobi–Bellman equation. J Optim Th Appl 96(3):589–626
2. Bellman R (1957) Dynamic programming. Princeton Univ. Press, Princeton
3. Bertsekas DP (1995) Dynamic programming and optimal control. Athena Sci., Belmont
4. Bryson AE, Ho Y (1975) Applied optimal control. Hemisphere, Washington, DC
5. Cahill AJ, James MR, Kieffer JC, Williamson D (1998) Remarks on the application of dynamic programming to the optimal path timing of robot manipulators. Internat J Robust and Nonlinear Control 8:463–482
6. Evans LC, James MR (1989) The Hamilton–Jacobi–Bellman equation for time-optimal control. SIAM J Control Optim 27(6):1477–1489
7. Sun M (1996) Alternating direction algorithms for solving Hamilton–Jacobi–Bellman equations. Applied Math Optim 34:267–277

Hemivariational Inequalities: Applications in Mechanics

EURIPIDIS MISTAKIDIS[1],
GEORGIOS E. STAVROULAKIS[2]
[1] University Thessaly, Volos, Greece
[2] Carolo Wilhelmina Techn. University, Braunschweig, Germany

MSC2000: 49S05, 74G99, 74H99, 74Pxx, 49J52, 90C33

Article Outline

Keywords

Hemivariational inequalities; Nonsmooth mechanics; Nonconvex energy function; Generalized subdifferential of F.H. Clarke

Variational expressions, also called for historical reasons *variational principles*, play a significant role in mechanics. They have their origin in the study of problems of analytical mechanics, which have extensively been studied in previous centuries, a time where scientists used to work multidisciplinary. Today, variational principles provide the basis for a correct and efficient modeling of a variety of physical phenomena, for instance, they provide the theoretical basis of the *finite element method* [19].

Variational equalities are the commonly met form of variational expressions. Having in mind problems which can be obtained from the minimization of a smooth (i. e., sufficiently differentiable) potential energy function, one may consider the variation of this function at a given point. A necessary condition for this function to attain a critical point is that every variation of the function in the neighborhood of this point is equal to zero. Thus, one formulates a variational equality problem. In mechanics, the differential of a potential energy function has the physical meaning of (stored or consumed) work. Let us consider a problem in *elastostatics*. In a formulation based on displacements, all variations of the system 's variables around a sought point are called *virtual displacements*. For obvious reasons the variational equality is called in this case *principle of virtual work*: for small virtual displacements around the equilibrium the virtual work of the system is equal to zero. Analogously, one arrives at the principles of complementary virtual work, or at mixed variational principles (the latter being derived from saddle point theorems). At this point it should be mentioned that a variational formulation may also be written for certain classes of problems which does not possess a potential.

The introduction of inequality constraints in the studied problem, or the assumption of nondifferentiable (nonsmooth) potential energy functions, lead to variational inequalities or more complicated variational problems. Intuitively speaking, either not all virtual variations of the problem variables around a given point are permitted (the case of inequality constraints, for instance, unilateral contact constraints), or, a linear approximation of the potential energy function is no more sufficient (the case of nondifferentiable or nonsmooth energy). Convex problems have certain theoretical and numerical advantages. They are connected with monotone operators. This is the case, e. g., of small displacement and deformation elastostatics with monotone material laws or interface and boundary conditions. These problems lead to *variational inequalities* and, in some cases, to convex (possibly nonsmooth) energy minimization problems (convex superpotentials in the sense of J.-J. Moreau [10]). The techniques of *convex analysis* and minimization can be used for their effective solution. Unilateral contact problems [10,15,17] and problems of elasto-

plasticity [7,17] have been studied within this framework.

Hemivariational inequalities are connected with nonconvex and possibly nonsmooth energy functions. In elastostatics, convexity is usually lost if the effects of large displacements or deformations are considered. Moreover, falling branches in material, interface or boundary laws lead to nonconvex potentials. The latter laws may be of a phenomenological nature and may be used for modeling of delamination and strength degradation effects, fracture, etc. Several methods have been developed for the study of nonconvex problems. The notion of the generalized gradient in the sense of F.H. Clarke has been used by P.D. Panagiotopoulos for the construction of hemivariational inequalities [16,17,18]. Following the example of nonsmooth analysis, he called this new field *nonsmooth mechanics*. A short introduction to this theory and its applications in mechanics is outlined in this article. The interested reader may also consult ▶ Nonconvex energy functions: Hemivariational inequalities and the monographs [14,18].

One should mention that the study of hemivariational inequalities provides an interesting field for mathematicians and engineers alike. For engineers several types of hemivariational inequalities have been used for the study and the efficient numerical treatment of yet unsolved or partially solved problems, e. g., in nonmonotone semipermeability problems, in modeling of delamination of simple and multilayered plates, in the theory of composite structures and adhesive joints, etc. Several of these concrete practical applications can not be treated by more naive, without mathematical justification engineering methods. Furthermore, the potential of this research field can be estimated if one thinks that nonconvex energy functions are connected with instabilities, complex dynamics, fractals and chaos. Certainly, a lot of work remains to be done in this area.

Abstract Hemivariational Inequality

The derivation of hemivariational inequalities is based on the mathematical notion of the generalized gradient of Clarke (denoted here by $\overline{\partial}$). In contrast to the variational inequalities, the hemivariational inequalities are not equivalent to minimum problems, but they give rise to substationarity problems. A hemivariational inequality problem reads: find $u \in V$ such as to satisfy the inequality

$$a(u, v-u) + \int_{\Omega} j^0(u, v-u)\, d\Omega \geq (l, v-u), \qquad \forall v \in V. \quad (1)$$

In the abstract form used here, let V be a real Hilbert space, V' be its dual space and such that $V \subset L^2(\Omega) \subset V'$, with continuous and dense injections. The problem is defined in Ω, which is an open bounded subset of $\mathbf{R}^n$. Furthermore let $(\cdot, \cdot)$ be the $L^2(\Omega)$ product and the duality pairing, $\| \cdot \|$ the norm of V and $| \cdot |_2$ the $L^2(\Omega)$-norm. Note that $(\cdot, \cdot)$ extends uniquely from $V \times L^2(\Omega)$ to $V \times V'$. Further, let $V \subset L^2(\Omega)$ be compact and $V \cap L^\infty(\Omega)$ be dense in V for the V-norm, and have a Galerkin base. The bilinear form $a(\cdot, \cdot)$: $V \times V \to \mathbf{R}$ is symmetric continuous and coercive, i. e. there exists $c > 0$ constant such that

$$a(v, v) \geq c \|v\|^2, \quad \forall v \in V. \quad (2)$$

Moreover j: $\mathbf{R} \to \mathbf{R}$ denotes a locally Lipschitz function which is defined by the following procedure: let $\beta \in L^\infty_{\text{loc}}(\mathbf{R})$ and consider

$$\overline{\beta}_\mu(\xi) = \text{esssup}_{|\xi_1 - \xi| \leq \mu} \beta(\xi_1) \quad (3)$$

and

$$\overline{\overline{\beta}}_\mu(\xi) = \text{essinf}_{|\xi_1 - \xi| \leq \mu} \beta(\xi_1). \quad (4)$$

They are increasing and decreasing functions of μ, respectively and thus the limits for $\mu \to 0_+$ exist. We denote them by $\overline{\beta}(\xi)$ and $\overline{\overline{\beta}}(\xi)$ respectively and we define the multivalued function

$$\widehat{\beta}(\xi) = [\overline{\beta}(\xi), \overline{\overline{\beta}}(\xi)]. \quad (5)$$

If $\beta(\xi_{\pm 0})$ exists for every $\xi \in \mathbf{R}$, then a locally Lipschitz function j: $\mathbf{R} \to \mathbf{R}$ can be determined (up to an additive constant) such that $\widehat{\beta}(\xi) = \overline{\partial} j(\xi)$. Finally, in relation (1) $j^0(u, v-u)$ denotes the generalized gradient of the nonconvex and nonsmooth locally Lipschitz potential j. By definition one has the following connection with the generalized gradient, in the sense of Clarke:

$$j^0(u, v) = \{\max \langle w, v \rangle : \ w \in \partial_{\text{CL}} j(u)\}. \quad (6)$$

Speaking in terms of mechanics one identifies relation (1) to be a virtual work expression in inequality form.

The first term is the internal work, the second term is the energy contribution of the nonlinear elements modeled by the nonconvex superpotential j and the right-hand side term represents the loading contribution. Detailed formulations of variational problems, up to the hemivariational inequality (1) for concrete applications follow in the next section.

Elastostatics with Nonlinear Boundary Conditions

A variational formulation is a statement that a solution of an operator equation subjected to certain boundary and/or initial conditions makes an expression involving variations of the quantities of the problems equal to zero or nonnegative. Thus one may distinguish between the bilateral or equality problems and the unilateral or inequality problems. Certain variational principles for a deformable body with nonlinear boundary interaction effects are derived in this section in order to demonstrate the hemivariational inequalities and their relation to classical equations and convex variational inequalities. Let $\Omega \in \mathbf{R}^3$ be an open bounded subset occupied by a deformable body in its undeformed state. On the assumption of small deformations we can write the relation:

$$\int_\Omega \sigma_{ij}(u)\varepsilon_{ij}(v-u)\,d\Omega = \int_\Omega f_i(v_i-u_i)\,d\Omega + \int_\Gamma \sigma_{ij}n_j(v_i-u_i)\,d\Gamma, \quad \forall v \in V, \quad (7)$$

for $u \in V$. Here V denotes the function space of the displacements which will be defined further. Relation (7) is the expression of the principle of virtual work for the body when it is considered free, without constraints on its boundary Γ. For the derivation of (7) the following steps are followed. The elastostatic equilibrium equation is first considered:

$$\sigma_{ij,j} + f_i = 0, \quad (8)$$

where the f_i is the volume force vector. Relation (8) is multiplied by the virtual variation $v_i - u_i$ and then an integration over Ω is performed. On the assumption of appropriately smooth functions, the Green – Gauss theorem is applied. One recalls here the strain-displacement relation (small deformation theory):

$$\varepsilon_{ij} = \frac{1}{2}(u_{i,j} + u_{j,i}). \quad (9)$$

Let a linearly elastic body be assumed, i. e., the constitutive material relation reads:

$$\sigma_{ij} = C_{ijhk}\varepsilon_{hk}, \quad (10)$$

where $C = \{ C_{ijhk} \}$, $i, j, h, k = 1, 2, 3$, is the elasticity tensor which satisfies the well-known symmetry and ellipticity properties

$$C_{ijhk} = C_{jihk} = C_{khij}, \quad (11)$$

$$C_{ijhk}\varepsilon_{ij}\varepsilon_{hk} \geq c\varepsilon_{ij}\varepsilon_{hk}, \quad \forall \varepsilon = \{\varepsilon_{ij}\}. \quad (12)$$

The bilinear form of linear elasticity $\alpha(\cdot,\cdot)$ reads in this case:

$$\alpha(u,v) = \int_\Omega C_{ijhk}\varepsilon_{ij}(u)\varepsilon_{hk}(v)\,d\Omega. \quad (13)$$

For further reference one splits the last term in (7) into the work of the normal and of the tangential tractions to the boundary. Then (7) may also be written in the form:

$$\int_\Omega \sigma_{ij}\varepsilon_{ij}(v-u)\,d\Omega = \int_\Omega f_i(v_i-u_i)\,d\Omega + \int_\Gamma S_N(v_N-u_N)\,d\Gamma + \int_\Gamma S_{T_i}(v_{T_i}-u_{T_i})\,d\Gamma, \quad \forall v \in V. \quad (14)$$

Single-Valued Boundary Laws and Variational Equalities

Let us assume first that on Γ the classical boundary conditions $S_N = 0$ and $u_{T_i} = 0$, $i = 1, 2, 3$, hold. Then (14) with (13) leads to the following variational equality:

$$\begin{cases} \text{Find} & u \in V_0 = \{v:\ v \in V,\ v_{T_i} = 0 \text{ on } \Gamma\} \\ \text{s.t.} & \alpha(u,v) = \int_\Omega f_i v_i\,d\Omega, \quad \forall v \in V_0. \end{cases} \quad (15)$$

Analogously, one treats all linear or nonlinear boundary conditions which can be expressed in an equality form. Relation (15), under appropriate smoothness assumptions, imply that the governing equations of the mechanical problem (8) and the assumed boundary conditions hold in a weak (integral or energetic) form.

Multivalued, Monotone Laws and Variational Inequalities

Let us assume now that on Γ the general monotone multivalued boundary condition

$$-S \in \partial j(u) \tag{16}$$

holds. Here $j(u)$ is assumed to be a convex superpotential and ∂denotes the subdifferential of convex analysis. Moreover, all (normal and tangential) contributions of boundary displacements u and tractions S are included in (16), which holds as a multidimensional boundary condition at each point of the boundary Γ. Relation (16) is, by definition of the subdifferential, equivalent to:

$$j(v) - j(u) \geq -S_i(v_i - u_i), \quad \forall v = \{v_i\} \in \mathbf{R}^3. \tag{17}$$

By using (17) and (7) one gets the variational inequality:

$$\begin{cases} \text{Find} & u \in V \text{ with } j(u) < \infty, \\ \text{s.t.} & \alpha(u, v-u) \\ & \quad + \displaystyle\int_\Gamma (j(v) - j(u))\, d\Gamma \\ & \quad \geq \displaystyle\int_\Omega f_i(v_i - u_i)\, d\Omega, \\ & \forall v \in V \text{ with } j(v) < \infty. \end{cases} \tag{18}$$

It is trivial to formulate analogous variational inequalities for more simple one-dimensional laws. This is the case where independent contact laws and tangential (e. g., due to friction) mechanisms are assumed on the boundary Γ. One should mention in passing that unilateral contact relations are included in this formulation by means of the indicator function in the place of $j(u)$. The *indicator function* is defined by $I_{U_{ad}}(u) = 0$ if $u \in U_{\text{ad}}$ and $+\infty$ otherwise, and includes the inequality constraints that describe the no-penetration requirements.

Multivalued, Nonmonotone Laws and Hemivariational Inequalities

In this case the basic building element is the definition of boundary conditions and material laws based on Clarke subdifferential (6). For instance, let on Γ the nonmonotone, possibly multivalued boundary condition

$$-S \in \partial_{\text{CL}} j(u) \tag{19}$$

hold, where j is a locally Lipschitz superpotential functional. Combining (7) with the inequality

$$j^0(u, v-u) \geq -S_i(v_i - u_i), \quad \forall v = \{v_i\} \in \mathbf{R}^3, \tag{20}$$

which defines on Γ the condition (19), one gets the following hemivariational inequality:

$$\begin{cases} \text{Find} & u \in V \\ \text{s.t.} & \alpha(u, v-u) + \\ & \quad + \displaystyle\int_\Gamma j^0(u, v-u)\, d\Gamma \\ & \quad \geq \displaystyle\int_\Omega f_i(v_i - u_i)\, d\Omega, \\ & \forall v \in V. \end{cases} \tag{21}$$

If instead of (19) one assumes on Γ that:

$$-S_N \in \partial_{\text{CL}} j_N(u_N), -S_T \in \partial_{\text{CL}} j_T(u_T), \tag{22}$$

then one gets analogously the hemivariational inequality:

$$\begin{cases} \text{Find} & u \in V \\ \text{s.t.} & \alpha(u, v-u) \\ & \quad + \displaystyle\int_\Gamma j_N^0(u_N, v_N - u_N)\, d\Gamma \\ & \quad + \displaystyle\int_\Gamma j_T^0(u_T, v_T - u_T)\, d\Gamma \\ & \quad \geq \displaystyle\int_\Omega f_i(v_i - u_i)\, d\Omega, \\ & \forall v \in V. \end{cases} \tag{23}$$

The last type of variational expressions involving $j^0(\cdot,\cdot)$ or $j_N^0(\cdot,\cdot)$ and $j_T^0(\cdot,\cdot)$ have been called *hemivariational inequalities* by Panagiotopoulos, who introduced and studied them in mechanics [14,16,17,18]. Note that in the more general case in which j or j_N and j_T are not locally Lipschitz $j^0(\cdot,\cdot)$ in (21) and $j_N^0(\cdot,\cdot)$, $j_T^0(\cdot,\cdot)$ in (23) are replaced by $j^\uparrow(\cdot,\cdot)$ and $j_N^\uparrow(\cdot,\cdot)$, $j_T^\uparrow(\cdot,\cdot)$. Moreover a combination of monotone subdifferential laws (cf. (6)) and nonmonotone laws (cf. (19)) for different (nonoverlapping) parts of the boundary Γ is possible. One then gets variational-hemivariational inequality problems.

The solution of variational problems, like the variational equalities, or the hemivariational inequalities derived previously, satisfies the operator equations of the problem, e. g. the equation of equilibrium, and the boundary conditions of the problem in a weak sense. This means, roughly speaking, that these relations are satisfied in an integral form, on the body or the boundary of the structure respectively. Analogous considerations are familiar within the weak formulations used in the finite element method.

Inequality or Nonsmooth Mechanics

A boundary value problem is called *bilateral* (resp. *unilateral*) if it leads to variational equality (resp. variational, or hemivariational inequality) formulations. The unilateral problems are called *inequality problems* too. Inequality problems in mechanics usually characterize structures with variable mechanical behavior, i. e. where the material or boundary law depends on the direction of the stress or boundary traction variation. Due to their connection with nonsmooth energy functions, all inequality problems belong to the area called by Panagiotopoulos *nonsmooth mechanics* [11,12].

Discretized Hemivariational Inequalities for Nonlinear Material Laws

In order to make the subject more accessible to engineers a discretized hemivariational inequality is formulated in this section. A finite element discretization is assumed. All relations are written in an elementary matrix analysis form. An elastic structure with both classical, linearly elastic and degrading elements is considered.

The *stress equilibrium equations* read:

$$\overline{G}\overline{s} = \begin{pmatrix} G & G_n \end{pmatrix} \begin{pmatrix} s \\ s_n \end{pmatrix} = p \tag{24}$$

where $\overline{G}$ is the equilibrium matrix of the discretized structure which takes into account the stress contribution of the linear s and nonlinear s_n elements and p is the loading vector.

The *strain-displacement compatibility equations* take the form:

$$\overline{e} = \begin{pmatrix} e \\ e_n \end{pmatrix} = \overline{G}^\top u = \begin{pmatrix} G^\top \\ G_n^\top \end{pmatrix} u, \tag{25}$$

where e, u are the deformation and displacement vectors respectively.

The linear material constitutive law for the structure reads:

$$s = K_0(e - e_0), \tag{26}$$

where K_0 is the natural and stiffness flexibility matrix and e_0 is the initial deformation vector.

The nonlinear material law is considered in the form:

$$s_n \in \partial_{\mathrm{CL}}\phi_n(e_n). \tag{27}$$

Here $\phi_n(\cdot)$, is a general nonconvex superpotential and summation over all nonlinear elements gives the total strain energy contribution of them as:

$$\Phi_n(e_n) = \sum_{i=1}^{q} \phi_n^{(i)}(e_n). \tag{28}$$

Finally classical support boundary conditions complete the description of the problem.

The discretized form of the virtual work equation reads:

$$s^\top(e^* - e) + s_n^\top(e_n^* - e_n) = p^\top(u^* - u), \quad \forall\, e^*,\ u^*,\ e_n^*. \tag{29}$$

Entering the elasticity law (26) into the virtual work equation (29), and using (25) we get:

$$u^\top G K_0^\top G^\top(u^* - u) - (p + GK_0e_0)^\top(u^* - u) + s_n^\top(e_n^* - e_n) = 0, \forall u^* \in V_{\mathrm{ad}}, \tag{30}$$

where $K = G^\top K_0^\top G$ denotes the stiffness matrix of the structure, $\overline{p} = p + GK_0 e_0$ denotes the nodal equivalent loading vector and V_{ad} includes all support boundary conditions of the structure.

Further one considers the nonlinear elements (27) in the inequality form:

$$s_n^\top (e_n^* - e_n) \le \Phi_n^o(e_n^* - e_n), \quad \forall e_n^*, \tag{31}$$

where $\Phi_n^o(e_n^* - e_n)$ is the directional derivative of the potential Φ_n. Thus the following discretized hemivariational inequality is obtained:

$$\begin{cases} \text{Find} & \text{kinematically admissible} \\ & \quad \text{displacements } u \in V_{\text{ad}} \\ \text{s.t.} & u^\top K(u^* - u) - \overline{p}^\top (u^* - u) \\ & \quad + \Phi_n^o(u_n^* - u_n) \ge 0, \\ & \forall u^* \in V_{\text{ad}}. \end{cases} \tag{32}$$

Equivalently a substationarity problem for the total potential energy can be written:

$$\begin{cases} \text{Find} & u \in V_{\text{ad}} \\ \text{s.t.} & \Pi(u) = \text{stat}_{v \in V_{\text{ad}}} \{\Pi(v)\}. \end{cases} \tag{33}$$

Here the potential energy reads $\Pi(v) = \frac{1}{2} v^\top K v - \overline{p}^\top v + \Phi_n(v)$, where the first two terms (quadratic potential) are well-known in the structural analysis community.

Other Applications in Mechanics

Hemivariational inequalities have been used for the modeling and solution of delamination effects in composite and multilayered plates, in composite structures, for nonmonotone friction and skin effects and for nonlinear mechanics applications (for instance, in the analysis of semi-rigid joints in steel structures). Details can be found in [9,11,12,17,18] and in the citations given there. Another area of applications are nonconvex problems arising in elastoplasticity (cf. [4,5,6]). Some nonconvex problems in elastoplasticity have been treated by hemivariational inequality techniques in [17,18]. Mathematical results which are useful for the study of hemivariational inequalities can also be found in [2,3,13,14].

Numerical Algorithms

A number of algorithms based on nonsmooth and nonconvex optimization concepts, on engineering methods or heuristics and on combination of these two approaches have been tested till now for the numerical solution of hemivariational inequality problems. Both finite elements and boundary elements have been used, the latter for boundary only nonlinear problems; see ▶ Nonconvex energy functions: Hemivariational inequalities and [1,8,9,17].

See also

- ▶ Generalized Monotonicity: Applications to Variational Inequalities and Equilibrium Problems
- ▶ Hemivariational Inequalities: Eigenvalue Problems
- ▶ Hemivariational Inequalities: Static Problems
- ▶ Nonconvex Energy Functions: Hemivariational Inequalities
- ▶ Nonconvex-nonsmooth Calculus of Variations
- ▶ Quasidifferentiable Optimization
- ▶ Quasidifferentiable Optimization: Algorithms for Hypodifferentiable Functions
- ▶ Quasidifferentiable Optimization: Algorithms for QD Functions
- ▶ Quasidifferentiable Optimization: Applications
- ▶ Quasidifferentiable Optimization: Applications to Thermoelasticity
- ▶ Quasidifferentiable Optimization: Calculus of Quasidifferentials
- ▶ Quasidifferentiable Optimization: Codifferentiable Functions
- ▶ Quasidifferentiable Optimization: Dini Derivatives, Clarke Derivatives
- ▶ Quasidifferentiable Optimization: Exact Penalty Methods
- ▶ Quasidifferentiable Optimization: Optimality Conditions
- ▶ Quasidifferentiable Optimization: Stability of Dynamic Systems
- ▶ Quasidifferentiable Optimization: Variational Formulations
- ▶ Quasivariational Inequalities
- ▶ Sensitivity Analysis of Variational Inequality Problems
- ▶ Solving Hemivariational Inequalities by Nonsmooth Optimization Methods

- ▶ Variational Inequalities
- ▶ Variational Inequalities: F. E. Approach
- ▶ Variational Inequalities: Geometric Interpretation, Existence and Uniqueness
- ▶ Variational Inequalities: Projected Dynamical System
- ▶ Variational Principles

References

1. Demyanov VF, Stavroulakis GE, Polyakova LN, Panagiotopoulos PD (1996) Quasidifferentiability and nonsmooth modelling in mechanics, engineering and economics. Kluwer, Dordrecht
2. Goeleven D (1996) Noncoercive variational problems and related results. Addison-Wesley and Longman
3. Haslinger J, Miettinen M, Panagiotopoulos PD (1999) Finite element method for hemivariational inequalities. Kluwer, Dordrecht
4. Kim SJ, Oden JT (1984) Generalized potentials in finite elastoplasticity. Part I. Internat J Eng Sci 22:1235–1257
5. Kim SJ, Oden JT (1985) Generalized potentials in finite elastoplasticity. Part II. Internat J Eng Sci 23:515–530
6. Kuczma MS, Stein E (1994) On nonconvex problems in the theory of plasticity. Arch Mechanicky 46(4):603–627
7. Maier G, Novati G (1990) Extremum theorems for finite-step backward-difference analysis of elastic-plastic nonlinearly hardening solids. Internat J Plasticity 6:1–10
8. Miettinen M, Mäkelä MM, Haslinger J (1995) On numerical solution of hemivariational inequalities by nonsmooth optimization methods. J Global Optim 8(4):401–425
9. Mistakidis ES, Stavroulakis GE (1998) Nonconvex optimization in mechanics. Algorithms, heuristics and engineering applications by the F.E.M. Kluwer, Dordrecht
10. Moreau JJ (1968) La notion de sur-potentiel et les liaisons unilatérales enélastostatique. CR 267A:954–957
11. Moreau JJ, Panagiotopoulos PD (eds) (1988) Nonsmooth mechanics and applications. CISM, vol 302. Springer, Berlin
12. Moreau JJ, Panagiotopoulos PD, Strang G (eds) (1988) Topics in nonsmooth mechanics. Birkhäuser, Basel
13. Motreanu D, Panagiotopoulos PD (1999) Minimax theorems and qualitative properties of the solutions of hemivariational inequalities. Kluwer, Dordrecht
14. Naniewicz Z, Panagiotopoulos PD (1995) Mathematical theory of hemivariational inequalities and applications. M. Dekker, New York
15. Oden JT, Kikuchi N (1988) Contact problems in elasticity: A study of variational inequalities and finite element methods. SIAM, Philadelphia
16. Panagiotopoulos PD (1983) Nonconvex energy functions. Hemivariational inequalities and substationary principles. Acta Mechanics 42:160–183
17. Panagiotopoulos PD (1985) Inequality problems in mechanics and applications. Convex and nonconvex energy functions. Birkhäuser, Basel
18. Panagiotopoulos PD (1993) Hemivariational inequalities. Applications in mechanics and engineering. Springer, Berlin
19. Washizu K (1968) Variational methods in elasticity and plasticity. Pergamon, Oxford

Hemivariational Inequalities: Eigenvalue Problems

Daniel Goeleven[1], Dumitru Motreanu[2]

[1] I.R.E.M.I.A., University de la Réunion, Saint-Denis, France

[2] Department Mat., University Al.I.Cuza, Iasi, Romania

MSC2000: 49J52

Article Outline

Keywords

Eigenvalue problem; Hemivariational inequalities; Critical point theory; Unilateral mechanics

The theory of *hemivariational inequalities* has been created by P.D. Panagiotopoulos et al. (see [3,5,6,7]) for studying nonconvex and nonsmooth energy functions under nonmonotone multivalued laws. In this setting many relevant models lead to *nonsmooth eigenvalue problems*. A typical example is provided by the analysis of hysteresis phenomena. To illustrate it we present here the loading and unloading problems with *hysteresis* modes.

Consider a plane linear elastic body Ω with the boundary Γ whose mechanical behavior is described by the virtual displacement variable u and the scalar parameter λ which determines the magnitude of the external loading on the system. The variable u must satisfy certain boundary or support conditions. For the sake of simplicity we assume that $u = 0$ on Γ, so the space of kinematically admissible displacements u is the *Sobolev*

space $H_0^1(\Omega)$, that is the closure of $C_0^\infty(\Omega)$ with respect to the L^2-norm of the gradient. Let us suppose that there exist a fundamental (pre-bifurcation) solution $\lambda \longmapsto u_0(\lambda)$ and another solution $\lambda \longmapsto u(\lambda) = u_0(\lambda) + z(\lambda)$ that coincide for $\lambda < \lambda_0$. Then one has $\lim_{\lambda\to\lambda_0} z(\lambda) = 0$ and the hysteresis bifurcation mode has the expression

$$u_1(\lambda_0) := \lim_{\lambda\to\lambda_0} \|z(\lambda)\|^{-1} z(\lambda). \tag{1}$$

Using the principle of virtual works together with physically realistic assumptions on the data θ and S (see e. g. [7]), we obtain the relation

$$a(u_1(\lambda_0), v) + \langle S(u_1(\lambda_0)), v\rangle - \lambda_0 \int_\Omega u_1(\lambda_0) v dx = 0, \quad \forall v \in H_0^1(\Omega). \tag{2}$$

It is justified to accept that a generalized nonmonotone reaction-displacement $(-S, u)$ holds in Ω expressed by the next law

$$\int_\Omega j^o(u_1(\lambda_0); v)\, dx \geq \langle S(u_1(\lambda_0)), v\rangle\,, \quad \forall v \in H_0^1(\Omega), \tag{3}$$

where j: $\mathbf{R} \to \mathbf{R}$ stands for a *locally Lipschitz function* with the *generalized gradient* ∂j and the *generalized directional derivative*

$$j^o(x; y) = \max\{\langle z, y\rangle : z \in \partial j(x)\}$$

(see [2]). Relations (2) and (3) yield the following eigenvalue problem in hemivariational inequality form: Find $(u = u(\lambda), \lambda) \in H_0^1(\Omega) \times \mathbf{R}$ such that

$$a(u, v) + \int_\Omega j^o(u; v)\, dx \geq \lambda \int_\Omega uv\, dx, \quad \forall v \in H_0^1(\Omega). \tag{4}$$

Additional information concerning problems of type (4) can be found in [3,5,6,7].

Relation (4), as well as other models, motivates the study of abstract eigenvalue problems for hemivariational inequalities. The specific case of Problem (4) can be reformulated as follows: given a Banach space V embedded in $L^2(\Omega)$, i. e. the space of square-integrable functions on $\Omega \subset \mathbf{R}^N$, a continuous symmetric bilinear form a: $V \times V \to \mathbf{R}$ and a locally Lipschitz function j: $\mathbf{R} \to \mathbf{R}$ with an appropriate growth condition for its generalized gradient, find $u \in V$ and $\lambda \in \mathbf{R}$ such that

$$a(u, v) + \int_\Omega j^o(u; v)\, dx \geq \lambda \int_\Omega uv dx, \quad \forall v \in V. \tag{5}$$

Note that this last mathematical model can also be used to formulate various other problems in Mechanics like unilateral bending problems in elasticity.

A general approach for studying the abstract eigenvalue problem (5) is the nonsmooth critical point theory as developed by K.-C. Chang [1]. In that paper the *minimax principles* in the critical point theory are extended from the smooth functionals (see [8]) to the case of locally Lipschitz functionals. In this respect we associate to Problem (5), for each λ, the locally Lipschitz functional I_λ: $V \to \mathbf{R}$,

$$I_\lambda(u) = \frac{1}{2} a(u, u) + \int_\Omega j(u)\, dx - \frac{\lambda}{2} \int_\Omega u^2\, dx, \quad \forall u \in V. \tag{6}$$

Note that a *critical point* u of I_λ, i. e. $0 \in \partial I_\lambda\,(u)$, is a solution of (5) because

$$\partial I_\lambda(u) \subset a(u, \cdot) - \lambda(u, \cdot)_{L^2} + \partial \int_\Omega j(u)\, dx \subset a(u, \cdot) - \lambda(u, \cdot)_{L^2} + \int_\Omega \partial j(u)\, dx$$

(see [2]). Thus, to solve (5), it suffices to establish the existence of nontrivial critical points of the functional I_λ introduced in (6). To this end we proceed along the lines in [4] by arguing in an abstract framework.

Given a Banach space V and a bounded domain Ω in $\mathbf{R}^m$, $m \geq 1$, let T: $V \to L^s(\Omega;\mathbf{R}^N)$ be a *compact linear operator*, where $L^s(\Omega;\mathbf{R}^N)$ stands for the Banach space of all Lebesgue measurable functions f: $\Omega \to \mathbf{R}^N$ for which $|f|^s$ is integrable with $1 < s < \infty$. Let F: $V \to \mathbf{R}$ be a locally Lipschitz function and let G: $\Omega \times \mathbf{R}^N \to \mathbf{R}$ be a (Carathéodory) function such that $G(x, y)$ is measurable in $x \in \Omega$, locally Lipschitz in $y \in \mathbf{R}^N$ and $G(x, 0) = F(0) = 0$, $x \in \Omega$. The hypotheses below are imposed

H1) $|w| \le c(1 + |y|^{s-1})$, $\forall w \in \partial_y G(x,y)$, $x \in \Omega$, $y \in \mathbf{R}^N$, with a constant $c > 0$;

H2) i) $F(v) - r\langle z, v_V\rangle \ge \alpha \|v\|_V^{\sigma} - \alpha_0$, $\forall v \in V$, $z \in \partial F(v)$;

ii) $G(x,y) - r\langle w, y\rangle \ge -b|y|^{\sigma_0} - b_0$, for a.e. $x \in \Omega$, $y \in \mathbf{R}^N$, $w \in \partial_y G(x,y)$, with positive constants $r, \alpha, \alpha_0, b, b_0, \sigma, \sigma_0$, where $1 \le \sigma_0 < \min\{\sigma, r^{-1}, s\}$;

H3) any bounded sequence $\{v_n\} \subset V$ for which there is $z_n \in \partial F(v_n)$ converging in V^* contains a convergent subsequence in V;

H4) i) $\liminf_{v\to 0} F(v)\|v\|_V^{-p} > 0$;

ii)

$$\liminf_{v\to 0} F(v)\|v\|_V^{-p} + |\Omega|^{(s-p)/p}\|T\|^p \liminf_{y\to 0} G(x,y)|y|^{-p} > 0$$

uniformly with respect to x, $1 \le p < s$;

H5)

$$\liminf_{t\to+\infty} F(tv_0)t^{-1/r} < -\liminf_{t\to+\infty} t^{-1/r}\int_\Omega G(x, tTv_0)\,dx$$

for some $v_0 \in V$.

The following statement is our main result in studying the abstract eigenvalue problem (5).

Theorem 1 *Assume that the hypotheses H1)–H5) hold. Then there exists a nontrivial critical point $u \in V$ of $I: V \to \mathbf{R}$ defined by*

$$I(v) = F(v) + \int_\Omega G(x,(Tv)(x))\,dx\,, \quad v \in V.$$

Moreover, there exists $z \in \partial F(u)$ and $w \in L^{s(s-1)}(\Omega;\mathbf{R}^N)$ such that

$$w(x) \in \partial_y G(x,(Tu)(x)) \quad \text{a.e. } x \in \Omega,$$
$$\langle z, v\rangle_V + \int_\Omega \langle w(x),(Tv)(x)\rangle\,dx = 0\,, \quad v \in V.$$

Conversely, if $u \in V$ verifies the relations above, corresponding to some z and w, and the function $G(x,\cdot)$ is regular at $(Tu)(x)$ (in the sense of F.H. Clarke [2]) for each $x \in \Omega$, then u is a critical point of I.

The foregoing locally Lipschitz functional I satisfies the Palais–Smale condition in the sense of Chang [1]. Indeed, let (v_n) be a sequence in V with $I(v_n) \le M$ and for which there exists a sequence $J_n \in \partial I(v_n)$ with $J_n \to 0$ in V^*. Then from H2) and taking into account that

$$J_n = z_n + T^* w_n,$$
$$z_n \in \partial F(v_n),$$
$$w_n(x) \in \partial_y G(x,(Tv_n)(x)) \quad \text{a.e. } x \in \Omega,$$

we infer that

$$M + r\|v_n\|_V \ge F(v_n) - r\langle z_n, v_n\rangle_V + \int_\Omega (G(x,(Tv_n)(x)) - r\langle w_n(x),(Tv_n)(x)\rangle)\,dx \ge \alpha\|v_n\|_V^{\sigma} + C_1\|v_n\|_V^{\sigma_0} + C_2,$$

with real constants C_1, C_2, provided that n is large enough. It is clear that the estimate above implies that the sequence (v_n) is bounded in V. Then a standard argument based on the assumption H3) allows to conclude that (v_n) possesses a strongly convergent subsequence. Namely, the boundedness of (v_n) implies that (Tv_n) is bounded in $L^s(\Omega;\mathbf{R}^N)$. Thus (w_n) is bounded in $L^{s/s-1)}(\Omega;\mathbf{R}^N)$ due essentially to the assumption H1). Since T^* is a compact operator and $J_n \to 0$ we derive that (z_n) has a convergent subsequence in V^*. This fact combined with the boundedness of (v_n) allows to use the hypothesis H3). The claim that the locally Lipschitz functional I verifies the Palais –Smale condition is proved.

Assumption H4) insures the existence of some constants $\delta > 0$, $A > 0$ and $B > 0$, with

$$A - B|\Omega|^{(s-p)/p}\|T\|^p > 0,$$

such that

$$F(v) \ge A\|v\|_V^p\,, \qquad \|v\|_V \le \delta, \tag{7}$$

and

$$G(x,y) \ge -B|y|^p\,, \quad \forall x \in \Omega, \quad |y| \le \delta.$$

Combining the inequality above with H1) one obtains that

$$\int_\Omega G(x,(Tv)(x))\,dx \ge -(A-\eta)\|v\|_V^p\,, \qquad \|v\|_V \le \rho, \tag{8}$$

for some $\eta > 0$ and $0 < \rho \le \delta$. Indeed, assumption H1) and Lebourg 's mean value theorem imply that G fulfills the following growth condition

$$|G(x,y)| \le a_1 + a_2 |y|^s, \forall x \in \Omega, \quad y \in \mathbf{R}^N,$$

with constants $a_1, a_2 \ge 0$. The two estimates above for $G(x,y)$ show that

$$G(x,y) \ge -B|y|^p - (a_1\delta^{-s} + a_2)|y|^s, \quad \forall x \in \Omega, \quad y \in \mathbf{R}^N.$$

Then one deduces from the continuity of T that one has

$$\int_\Omega G(x,(Tv)(x))dx \ge \left(-B|\Omega|^{(s-p)/p}\|T\|^p - (a_1\delta^{-s} + a_2)\|T\|^s \|v\|_V^{s-p}\right)\|v\|_V^p, \quad \forall v \in V.$$

Since $s > p$ we see that the numbers $\eta > 0$ and $\rho > 0$ can be chosen so small that relation (8) be verified.

By (7) and (8) we arrive at the conclusion that there exist positive numbers ρ, η such that

$$I(v) \ge \eta, \quad \|v\|_V = \rho. \tag{9}$$

The formula

$$\partial_t(t^{-1/r}G(x,ty)) = \frac{1}{r}t^{-1-1/r}[r\langle \partial_y G(x,ty), ty\rangle - G(x,ty)],$$

the absolute continuity property and H2ii) show that

$$t^{-1/r}G(x,ty) - G(x,y) = \int_1^t \partial_\tau(\tau^{-1/r}(G(x,\tau y)d\tau \le C|y|^{\sigma_0} + C_0$$

for a.e. $x \in \Omega$, $y \in \mathbf{R}^N$, $t > 1$, where C, C_0 are positive constants. Then one obtains

$$I(t\theta v_0) \le (t\theta)^{1/r} \times \left[F(t\theta v_0)(t\theta)^{-1/r} + \overline{C}\|v_0\|_V^{\sigma_0}\theta^{\sigma_0 - 1/r} + \overline{C}_0\theta^{-1/r} + \theta^{-1/r}\int_\Omega G(x,\theta(Tv_0)(x))\,dx\right]$$

for all $t > 1$, $\theta > 1$, with new positive constants $\overline{C}$, $\overline{C}_0$. In view of H5) and since $\sigma_0 < 1/r$, we can find θ sufficiently large such that

$$\overline{C}\|v_0\|_V^{\sigma_0}\theta^{\sigma_0 - 1/r} + \overline{C}_0\theta^{-1/r} + \theta^{-1/r}\int_\Omega G(x,\theta(Tv_0)(x))\,dx < -\liminf_{\tau\to+\infty} F(\tau v_0)\tau^{-1/r}.$$

With such fixed number θ, we see that there exists arbitrarily large t satisfying

$$F(t\theta v_0)(t\theta)^{-1/r} + \overline{C}\|v_0\|_V^{\sigma_0}\theta^{\sigma_0 - 1/r} + \overline{C}_0\theta^{-1/r} + \theta^{-1/r}\int_\Omega G(x,\theta(Tv_0)(x))\,dx < 0.$$

We deduce that

$$I(t_n v_0) \le 0 \tag{10}$$

for a subsequence $t_n \to \infty$. The properties (9) and (10) permit to apply the *mountain pass theorem* in the nonsmooth version of Chang [1]. This yields the desired critical point u of I. The other assertions of the first part of Theorem are direct consequences of the last statement.

The converse part of Theorem follows from the next formula

$$\partial\int_\Omega G(x,u(x))\,dx = \int_\Omega \partial_y G(x,u(x))\,dx, \quad \forall u \in L^s(\Omega;\mathbf{R}^N),$$

which is valid under the growth condition in H1) and the regularity assumption for G (see [2]). The proof of Theorem is thus complete.

In the case of problem (4) we choose $V = H_0^1(\Omega)$, the compact linear operator T: $H_0^1(\Omega) \to L^s(\Omega)$ equal to the embedding $H_0^1(\Omega) \subset L^s(\Omega)$ with $2 < s < 2m(m-2)^{-1}$ if $m \ge 3$,

$$F(v) = \frac{1}{2}\int_\Omega (|\nabla v|^2 - \lambda v^2)\,dx, \quad \forall v \in H_0^1(\Omega),$$

where for simplicity we take $a(u,v) = \int_\Omega \nabla u \cdot \nabla v\,dx$, and $G(x,t) = j(t)$. A significant possible choice for j is the following one

$$j(t) = -\frac{|t|^s}{s} + \int_0^t \beta(\tau)\,d\tau, \quad t \in \mathbf{R}, \tag{11}$$

where $\beta \in L^\infty_{loc}(\mathbf{R})$ verifies $t\,\beta(t) \ge 0$ for t near 0, $|\beta(t)| \le c(1 + |t|^\gamma)$, $t \in \mathbf{R}$, with constants $c > 0$, $0 \le \gamma < 1$.

Corollary 2 *Let $j: \boldsymbol{R} \to \boldsymbol{R}$ be given by (11). If λ_1 denotes the first eigenvalue of $-\Delta$ on $H_0^1(\Omega)$, then for every $\lambda < \lambda_1$ the problem (5) with a as above, has a nontrivial eigenfunction $u \in H_0^1(\Omega)$ which solves in addition the nonsmooth Dirichlet problem containing both superlinear and sublinear terms*

$$\Delta u + \lambda u + |u|^{s-2} u \in [\underline{\beta}(u(x)), \overline{\beta}(u(x))]$$
$$\text{a.e. } x \in \Omega, u = 0 \quad \text{on } \partial\Omega,$$

where the notations in [1] are used.

The argument consists in verifying the assumptions H1)–H5) for the functional $I = I_\lambda$, for $\lambda < \lambda_1$, with I_λ described in (6). To this end it is sufficient to take $r \in (1/s, 1/2)$, $p = \sigma = 2$, $\sigma_0 = \gamma + 1$ and $v_0 \in H_0^1(\Omega) \{ 0 \}$. Applying Theorem one finds the stated result.

Other related results and applications for eigenvalue problems in the form of hemivariational inequalities are given in [3,4,5,6,7] and the references therein.

See also

- ▶ αBB Algorithm
- ▶ Eigenvalue Enclosures for Ordinary Differential Equations
- ▶ Generalized Monotonicity: Applications to Variational Inequalities and Equilibrium Problems
- ▶ Hemivariational Inequalities: Applications in Mechanics
- ▶ Hemivariational Inequalities: Static Problems
- ▶ Interval Analysis: Eigenvalue Bounds of Interval Matrices
- ▶ Nonconvex Energy Functions: Hemivariational Inequalities
- ▶ Nonconvex-nonsmooth Calculus of Variations
- ▶ Quasidifferentiable Optimization
- ▶ Quasidifferentiable Optimization: Algorithms for Hypodifferentiable Functions
- ▶ Quasidifferentiable Optimization: Algorithms for QD Functions
- ▶ Quasidifferentiable Optimization: Applications
- ▶ Quasidifferentiable Optimization: Applications to Thermoelasticity
- ▶ Quasidifferentiable Optimization: Calculus of Quasidifferentials
- ▶ Quasidifferentiable Optimization: Codifferentiable Functions
- ▶ Quasidifferentiable Optimization: Dini Derivatives, Clarke Derivatives
- ▶ Quasidifferentiable Optimization: Exact Penalty Methods
- ▶ Quasidifferentiable Optimization: Optimality Conditions
- ▶ Quasidifferentiable Optimization: Stability of Dynamic Systems
- ▶ Quasidifferentiable Optimization: Variational Formulations
- ▶ Quasivariational Inequalities
- ▶ Semidefinite Programming and Determinant Maximization
- ▶ Sensitivity Analysis of Variational Inequality Problems
- ▶ Solving Hemivariational Inequalities by Nonsmooth Optimization Methods
- ▶ Variational Inequalities
- ▶ Variational Inequalities: F. E. Approach
- ▶ Variational Inequalities: Geometric Interpretation, Existence and Uniqueness
- ▶ Variational Inequalities: Projected Dynamical System
- ▶ Variational Principles

References

1. Chang K-C (1981) Variational methods for non-differentiable functionals and their applications to partial differential equations. J Math Anal Appl 80:102–129
2. Clarke FH (1984) Nonsmooth analysis and optimization. Wiley, New York
3. Goeleven D, Motreanu D, Panagiotopoulos PD (1997) Multiple solutions for a class of eigenvalue problems in hemivariational inequalities. Nonlinear Anal Th Methods Appl 29:9–26
4. Motreanu D (1995) Existence of critical points in a general setting. Set-Valued Anal, 3:295–305
5. Motreanu D, Panagiotopoulos PD (1999) Minimax theorems and qualitative properties of the solutions of hemivariational inequalities. Kluwer, Dordrecht
6. Naniewicz Z, Panagiotopoulos PD (1995) The mathematical theory of hemivariational inequalities and applications. M. Dekker, New York
7. Panagiotopoulos PD (1993) Hemivariational inequalities. applications in mechanics and engineering. Springer, Berlin
8. Rabinowitz PH (1986) Minimax methods in critical point theory with applications to differential equations, vol 65. CBMS Reg. Conf. Ser. Math., Amer. Math. Soc., Providence

Hemivariational Inequalities: Static Problems

HVI

Zdzisław Naniewicz[1,2]
[1] Institute Appl. Math. Mech., Warsaw University, Warsaw, Poland
[2] Institute Math. Comp. Science, Techn. University Czestochowa, Czestochowa, Poland

MSC2000: 49J40, 47J20, 49J40, 35A15

Article Outline

Keywords

Semicoercive hemivariational inequality; Unilateral growth condition; Pseudomonotone mapping; Recession functional

Let $V = H^1(\Omega;\mathbf{R}^N)$, $N \geq 1$, be a vector valued Sobolev space of functions square integrable together with their first partial distributional derivatives in Ω, Ω being a bounded domain in $\mathbf{R}^m$, $m > 2$, with sufficiently smooth boundary Γ. Assume that V is compactly imbedded into $L^p(\Omega;\mathbf{R}^N)$ $(1 < p < 2m/m - 2)$, [12]). We write $\| \cdot \|_V$ and $\| \cdot \|_{L^p(\Omega;R^N)}$ for the norms in V and $L^p(\Omega;\mathbf{R}^N)$, respectively. For the pairing over $V^* \times V$ the symbol $\langle \cdot, \cdot \rangle_V$ will be used, V^* being the dual of V.

Let A: $V \to V^*$ be a bounded, *pseudomonotone operator*. This means that A maps bounded sets into bounded sets and that the following conditions hold [3,5]:

i) The effective domain of A coincides with the whole V;

ii) If $u_n \to u$ weakly in V and $\limsup_{n\to\infty} \langle Au_n, u_n - u_V \rangle \leq 0$, then $\liminf_{n\to\infty} \langle Au_n, u_n - v_V \rangle \geq \langle Au, u - v \rangle_V$ for any $v \in V$.
Note that i) and ii) imply that A is demicontinuous, i. e.

iii) If $u_n \to u$ strongly in V, then $Au_n \to Au$ weakly in V^*.

Moreover, we assume that V is endowed with a direct sum decomposition $V = \widehat{V} + V_0$, where V_0 is a finite-dimensional linear subspace, with respect to which A is semicoercive, i. e. $\forall u \in V$ there exist $\widehat{u} \in \widehat{V}$ and $\theta \in V_0$ such that $u = \widehat{u} + \theta$ and

$$\langle Au, u \rangle_V \geq c(\|\widehat{u}\|_V)\, \|\widehat{u}\|_V\,, \tag{1}$$

where c: $\mathbf{R}^+ \to \mathbf{R}$ stands for a coercivity function with $c(r) \to \infty$ as $r \to \infty$. Further, let j: $\mathbf{R}^N \to \mathbf{R}$ be a locally Lipschitz function fulfilling the *unilateral growth conditions* ([16,21]):

$$j^0(\xi;\eta - \xi) \leq \alpha(r)(1 + |\xi|^\sigma), \qquad \forall \xi, \eta \in \mathbf{R}^N,\ |\eta| \leq r,\ r \geq 0, \tag{2}$$

and

$$j^0(\xi;-\xi) \leq k\,|\xi|\,, \quad \forall \xi \in \mathbf{R}^N, \tag{3}$$

where $1 \leq \sigma < p$, k is a nonnegative constant and α :$\mathbf{R}^+ \to \mathbf{R}^+$ is assumed to be a nondecreasing function from $\mathbf{R}^+$ into $\mathbf{R}^+$. Here, $j^0(\cdot;\cdot)$ stands for the *directional Clarke derivative*

$$j^0(\xi;\eta) = \limsup_{\substack{h\to 0\\ \lambda \to 0_+}} \frac{j(\xi + h + \lambda\eta) - j(\xi + h)}{\lambda}, \tag{4}$$

by means of which the *Clarke generalized gradient* of j is defined by [6]

$$\partial j(\xi) := \{\mu \in \mathbf{R}^N :\ j^0(\xi;\eta) \geq \mu \cdot \eta,\ \forall \eta \in \mathbf{R}^N\}\,, \qquad \xi, \eta \in \mathbf{R}^N.$$

Remark 1 The unilateral growth condition (2) is the generalization of the well known sign condition used for the study of nonlinear partial differential equations in the case of scalar-valued function spaces (cf. [27,28]).

Consider the problem of finding $u \in V$ such as to satisfy the hemivariational inequality

$$\langle Au - g, v - u \rangle_V + \int_\Omega j^0(u; v - u)\, d\Omega \geq 0, \qquad \forall v \in V. \tag{5}$$

It will be assumed that $g \in V^*$ fulfills the *compatibility condition*

$$\langle g, \theta \rangle_V < \int_\Omega j^\infty(\theta)\, d\Omega, \quad \forall \theta \in V_0 \setminus \{0\}, \tag{6}$$

where j^∞: $\mathbf{R}^N \to \mathbf{R} \cup \{+\infty\}$ stands for the *recession functional* given by (cf. [2,4,10])

$$j^\infty(\xi) = \liminf_{\substack{\eta\to\xi \\ t\to+\infty}} [-j^0(t\eta;-\eta)], \quad \xi \in \mathbf{R}^N. \tag{7}$$

Because of (1), the problem to be considered here will be referred to as a *semicoercive hemivariational inequality*.

The notion of hemivariational inequality has been first introduced by P.D. Panagiotopoulos in [22,23] for the description of important problems in physics and engineering, where nonmonotone, multivalued boundary or interface conditions occur, or where some nonmonotone, multivalued relations between stress and strain, or reaction and displacement have to be taken into account. The theory of hemivariational inequalities (as the generalization of variational inequalities, cf. [7]) has been proved to be very useful in understanding of many problems of mechanics involving nonconvex, nonsmooth energy functionals. For the general study of hemivariational inequalities and their applications, see [13,14,15,17,18,19,20,21,24,26] and the references quoted there. Some results in the area of static, semicoercive inequality problems can be found in [9,10,25].

To prove the existence of solutions to (5), the Galerkin method combined with the pseudomonotone regularization of the nonlinearities will be applied.

Let us start with the following preliminary results.

The regularization $\widetilde{j}^0_R(\cdot;\cdot)$, $R>0$, of the Clarke directional derivative $j^0(\cdot;\cdot)$ will be defined as follows: for any $\xi, \eta \in \mathbf{R}^N$, set

$$\widetilde{j}^0_R(\xi,\eta) = \begin{cases} j^0(\xi;\eta) & \text{if } |\xi| \le R, \\ j^0\left(R\frac{\xi}{|\xi|};\eta\right) & \text{if } |\xi| > R. \end{cases} \tag{8}$$

Lemma 2 *Suppose that (2) and (3) are fulfilled. Then for $R>0$,*

$$\widetilde{j}^0_R(\xi;\eta-\xi) \le \widetilde{\alpha}(r)(1+|\xi|^\sigma), \quad \forall \xi \in \mathbf{R}^N, \quad \forall \eta \in \mathbf{R}^N,\ |\eta| \le r,\ r \ge 0. \tag{9}$$

$$\widetilde{j}^0_R(\xi;-\xi) \le k|\xi|, \quad \forall \xi \in \mathbf{R}^N, \tag{10}$$

where $\widetilde{\alpha}$: $\mathbf{R}^+ \to \mathbf{R}^+$ is a nondecreasing function independent of R.

Proof To establish (9) and (10) it suffices to consider the case $|\xi| \ge R$ and to invoke the estimates

$$\begin{aligned} \widetilde{j}^0_R(\xi;\eta-\xi) &= j^0\left(R\frac{\xi}{|\xi|};\eta-\xi\right) \\ &\le j^0\left(R\frac{\xi}{|\xi|};\eta-R\frac{\xi}{|\xi|}\right) \\ &\quad + \frac{|\xi|-R}{R} j^0\left(R\frac{\xi}{|\xi|};-R\frac{\xi}{|\xi|}\right) \\ &\le \alpha(|\eta|)(1+R^\sigma) + \frac{|\xi|-R}{R}kR \\ &\le \alpha(r)(1+|\xi|^\sigma) + k|\xi|, \\ &\forall \xi,\eta \in \mathbf{R}^N,\ |\eta| \le r,\ r \ge 0, \end{aligned}$$

and

$$\begin{aligned} \widetilde{j}^0_R(\xi;-\xi) &= j^0\left(R\frac{\xi}{|\xi|};-\xi\right) \\ &\le \frac{|\xi|}{R} j^0\left(R\frac{\xi}{|\xi|};-R\frac{\xi}{|\xi|}\right) \le \frac{|\xi|}{R}kR = k|\xi|, \end{aligned}$$

respectively. The proof is complete.

For any $R>0$, the following regularization of the primal problem can be formulated:

(P_R) Find $(u_R, \chi_R) \in V \times L^q(\Omega;\mathbf{R}^N)$,

1/p +1/q = 1, such that

$$\langle Au_R - g, v-u_R\rangle_V + \int_\Omega \chi_R \cdot (v-u_R)\,d\Omega = 0, \quad \forall v \in V, \tag{11}$$

$$\chi_R \in \Gamma_R(u_R), \tag{12}$$

where

$$\Gamma_R(u_R) := \left\{ \psi \in L^q(\Omega;\mathbf{R}^N) : \int_\Omega \psi \cdot v\,d\Omega \le \int_\Omega \widetilde{j}^0_R(u_R;v)d\Omega,\ \forall\, v \in L^p(\Omega;\mathbf{R}^N) \right\}.$$

In order to show that (P_R) has solutions, the following auxiliary result is to be applied.

Lemma 3 *Suppose that (1)-(3) and (6) hold. Then there exists $R_0 > 0$ such that for any $R > R_0$ the set of all $u \in V$ with the property that*

$$\langle Au - g, u\rangle_V - \int_\Omega \widetilde{j}_R^0(u; -u)\, d\Omega \le 0 \tag{13}$$

is bounded in V, i. e. there exists $\mathcal{M} > 0$ (possibly depending on $R > R_0$), such that (13) implies

$$\|u\|_V \le \mathcal{M}. \tag{14}$$

Proof Suppose on the contrary that this claim is not true, i. e. there exists a sequence $\{u_n\}_{n=1}^\infty \subset V$ with the property that

$$\langle Au_n - g, u_n\rangle_V - \int_\Omega \widetilde{j}_R^0(u_n; -u_n)\, d\Omega \le 0, \tag{15}$$

where $\|u_n\|_V \to \infty$ as $n \to \infty$. By the hypothesis, each element u_n can be represented as

$$u_n = \widehat{u}_n + e_n\theta_n, \tag{16}$$

where $\widehat{u}_n \in \widehat{V}$, $e_n \ge 0$, $\theta_n \in V_0$, $\|\theta_n\|_V = 1$, and $\langle Au_n, u_n\rangle_V \ge c(\|\widehat{u}_n\|_V)\|\widehat{u}_n\|_V$). Taking into account (3) it follows that

$$\begin{aligned}
0 &\ge \langle Au_n - g, u_n\rangle_V - \int_\Omega \widetilde{j}_R^0(u_n; -u_n)\, d\Omega \\
&\ge c(\|\widehat{u}_n\|_V)\|\widehat{u}_n\|_V - \|g\|_{V^*}\|\widehat{u}_n\|_V \\
&\quad - e_n\langle g, \theta_n\rangle_V - k\int_\Omega |u_n|\, d\Omega \\
&\ge c(\|\widehat{u}_n\|_V)\|\widehat{u}_n\|_V - \|g\|_{V^*}(\|\widehat{u}_n\|_V + e_n) \\
&\quad - k_1\|\widehat{u}_n\|_V - e_n k_1\|\theta_n\|_V,
\end{aligned} \tag{17}$$

where $k_1 = \text{const}$. The obtained estimates imply that $\{e_n\}$ is unbounded. Indeed, if it would not be so, then due to the behavior of $c(\cdot)$ at infinity, $\{\widehat{u}_n\}$ had to be bounded. In such a case the contradiction with $\|u_n\|_V \to \infty$ as $n \to \infty$ results. Therefore one can suppose without loss of generality that $e_n \to +\infty$ as $n \to \infty$. The next claim is that

$$\frac{1}{e_n}\widehat{u}_n \to 0 \quad \text{strongly in } V. \tag{18}$$

Indeed, if $\{\|\widehat{u}_n\|_V\}$ is bounded, then (18) follows immediately. If $\|\widehat{u}_n\|_V \to \infty$ then $c(\|\widehat{u}_n\|_V) \to +\infty$. From (17) one has

$$k_1 + \|g\|_{V^*} \ge \left(c(\|\widehat{u}_n\|_V) - \|g\|_{V^*} - k_1\right)\frac{\|\widehat{u}_n\|_V}{e_n}.$$

Thus, the boundedness of the sequence

$$\left\{\left(c(\|\widehat{u}_n\|_V) - \|g\|_{V^*} - k_1\right)\frac{\|\widehat{u}_n\|_V}{e_n}\right\}_{n=1}^\infty$$

results, which in view of

$$c(\|\widehat{u}_n\|_V) - \|g\|_{V^*} - k_1 \to +\infty \quad \text{as } n \to \infty$$

implies the assertion (18). The obtained results give rise to the following representation of u_n:

$$u_n = e_n\left(\frac{1}{e_n}\widehat{u}_n + \theta_n\right),$$

where $\widehat{u}_n/e_n \to 0$ strongly in V and $\theta_n \to \theta$ in V_0 as $n \to \infty$ for some $\theta \in V_0$ with $\|\theta\|_V = 1$ (recall that V_0 has been assumed to be finite dimensional). Moreover, the compact imbedding $V \subset L^p(\Omega;\mathbf{R}^N)$ permits one to suppose that $\widehat{u}_n/e_n \to 0$ and $\theta_n \to \theta$ a.e. in Ω.

Further, (15), together with the fact that A is semicoercive, leads to

$$\begin{aligned}
0 &\ge \langle Au_n - g, u_n\rangle_V - \int_\Omega \widetilde{j}_R^0(u_n; -u_n)d\Omega \\
&\ge \left(c(\|\widehat{u}_n\|_V) - \|g\|_{V^*}\right)\|\widehat{u}_n\|_V - e_n\langle g, \theta_n\rangle_V \\
&\quad + e_n \\
&\quad \cdot \int_\Omega -\widetilde{j}_R^0\left(e_n\left(\frac{1}{e_n}\widehat{u}_n + \theta_n\right); -\frac{1}{e_n}\widehat{u}_n - \theta_n\right) d\Omega.
\end{aligned}$$

Hence

$$\begin{aligned}
\langle g, \theta_n\rangle_V &\ge \left(c(\|\widehat{u}_n\|_V) - \|g\|_{V^*}\right)\frac{1}{e_n}\|\widehat{u}_n\|_V \\
&\quad + \int_\Omega -\widetilde{j}_R^0\left(e_n\left(\frac{\widehat{u}_n}{e_n} + \theta_n\right); -\frac{\widehat{u}_n}{e_n} - \theta_n\right) d\Omega.
\end{aligned} \tag{19}$$

Now observe that either

$$\left(c(\|\widehat{u}_n\|_V) - \|g\|_{V^*}\right)\frac{1}{e_n}\|\widehat{u}_n\|_V \to 0 \quad \text{as } n \to \infty,$$

if $\{\|\widehat{u}_n\|_V\}$ is bounded, or

$$\left(c(\|\widehat{u}_n\|_V) - \|g\|_{V^*}\right)\frac{1}{e_n}\|\widehat{u}_n\|_V \geq 0$$

for sufficiently large n, if $\|\widehat{u}_n\|_V \to \infty$ as $n \to \infty$. Therefore, for any case

$$\liminf_{n\to\infty}\left(c(\|\widehat{u}_n\|_V) - \|g\|_{V^*}\right)\frac{1}{e_n}\|\widehat{u}_n\|_V \geq 0.$$

Moreover, by (10) the estimate follows:

$$-\widetilde{j}_R^0\left(e_n\left(\frac{\widehat{u}_n}{e_n}+\theta_n\right);-\frac{\widehat{u}_n}{e_n}-\theta_n\right) \geq -k\left|\frac{\widehat{u}_n}{e_n}+\theta_n\right|. \quad (20)$$

This allows the application of Fatou 's lemma in (19), from which one is led to

$$\begin{aligned}\langle g,\theta\rangle_V &\geq \liminf_{n\to\infty}\int_\Omega\left[-\widetilde{j}_R^0\left(e_n\left(\frac{\widehat{u}_n}{e_n}+\theta_n\right);-\frac{\widehat{u}_n}{e_n}-\theta_n\right)\right]d\Omega\\ &\geq \int_\Omega \liminf_{n\to\infty}\left[-\widetilde{j}_R^0\left(e_n\left(\frac{\widehat{u}_n}{e_n}+\theta_n\right);-\frac{\widehat{u}_n}{e_n}-\theta_n\right)\right]d\Omega.\end{aligned} \quad (21)$$

Taking into account (8) and upper semicontinuity of $j^0(\cdot,\cdot)$, one can easily verify that

$$\begin{aligned}&\liminf_{n\to\infty}\left[-\widetilde{j}_R^0\left(e_n(\frac{\widehat{u}_n}{e_n}+\theta_n);-\frac{\widehat{u}_n}{e_n}-\theta_n\right)\right]\\ &\geq -j^0\left(R\frac{\theta}{|\theta|};-\theta\right),\end{aligned}$$

which leads to

$$\langle g,\theta\rangle_V \geq \int_\Omega -j^0\left(R\frac{\theta}{|\theta|};-\theta\right)d\Omega. \quad (22)$$

Since $j^\infty(\cdot)$ is lower semicontinuous and V_0 is finite dimensional, from (6) it follows that a $\delta > 0$ can be found such that for any $\theta \in V_0$ with $\|\theta\|_V\| = 1$,

$$\langle g,\theta\rangle_V + \delta < \int_\Omega j^\infty(\theta)\,d\Omega. \quad (23)$$

With the help of Fatou 's lemma (permitted by (20)) we arrive at

$$\liminf_{R\to\infty}\int_\Omega -j^0\left(\frac{R}{|\theta|}\theta;-\theta\right)d\Omega \geq \int_\Omega j^\infty(\theta)\,d\Omega.$$

The upper semicontinuity of $j^0(\cdot;\cdot)$ allows us to conclude the existence of $R_\theta > 0$ and $\varepsilon_\theta > 0$ such that

$$\int_\Omega -j^0\left(\frac{R}{|\theta'|}\theta';-\theta'\right)d\Omega \geq \int_\Omega j^\infty(\theta)\,d\Omega - \frac{\delta}{2}$$

for each $R > R_\theta$ and $\theta' \in V_0$ with $\|\theta - \theta'\|_V\| < \varepsilon_\theta$. As the sphere $\{v \in V_0 : \|v\|_V\| = 1\}$ is compact in V_0, there exists $R_0 > 0$ such that

$$\int_\Omega -j^0\left(\frac{R}{|\theta|}\theta;-\theta\right)d\Omega \geq \int_\Omega j^\infty(\theta)\,d\Omega - \frac{\delta}{2},$$

for any $\theta \in V_0$ with $\|\theta\|_V\| = 1$, $R > R_0$. This combined with (23) contradicts (22). Accordingly, the existence of a constant $\mathcal{M} > 0$ has been established such that (13) implies (14), whenever $R > R_0$. The proof of Lemma 3 is complete.

Proposition 4 *Let us assume all the hypotheses stated above. Then for any $R > R_0$ the problem (P_R) possesses at least one solution. Moreover, if (u_R, χ_R) is a solution of (P_R), then*

$$\|u_R\|_V\| \leq M \quad (24)$$

for some constant M not depending on $R > R_0$.

Proof Let Λ be the family of all finite-dimensional subspaces F of V, ordered by inclusion. Denote by i_F: $F \to V$ the inclusion mapping of F into V and by i_F^*: $V^* \to F^*$ the dual projection mapping of V^* into F^*, F^* being the dual of F. The pairing over $F^* \times F$ will be denoted by $\langle\cdot,\cdot\rangle_F$. Set $A_F := i_F^* \circ A \circ i_F$ and $g_F := i_F^* g$.

Fix $R > R_0$. For any $F \in \Lambda$ consider a finite-dimensional regularization of (P_R):

(P_F) Find $(u_F, \chi_F) \in F \times L^q(\Omega;\mathbf{R}^N) \in F$

such that

$$\langle Au_F - g, v\rangle_V + \int_\Omega \chi_F \cdot v\,d\Omega = 0, \forall v \in F, \quad (25)$$

$$\chi_F \in \Gamma_R(u_F). \quad (26)$$

The first task is to show that for each $F \in \Lambda$, (P_F) has solutions. Notice that $\Gamma_R(\cdot)$ has nonempty, convex and closed values and if $\psi \in \Gamma_R(v)$, $v \in L^p(\Omega;\mathbf{R}^N)$, then

$$\|\psi\|_{L^q(\Omega;\mathbf{R}^N)} \leq K_R, \tag{27}$$

for some $K_R > 0$ depending on the Lipschitz constant of j in the ball $\{\eta \in \mathbf{R}^N : |\eta| \leq \mathbf{R}\}$. Moreover, from the upper semicontinuity of $j_R^0(\cdot;\cdot)$ and Fatou 's lemma it follows immediately that Γ_R is upper semicontinuous from $L^p(\Omega;\mathbf{R}^N)$ to $L^q(\Omega;\mathbf{R}^N)$, $L^q(\Omega;\mathbf{R}^N)$ being endowed with the weak topology.

Further, let $\tau_F: L^q(\Omega;\mathbf{R}^N) \to F^*$ be the operator that to any $\psi \in L^q(\Omega;\mathbf{R}^N)$ assigns $\tau_F \psi \in F^*$ defined by

$$\langle \tau_F \psi, v\rangle_F := \int_\Omega \psi \cdot v \, d\Omega \quad \text{for any } v \in F. \tag{28}$$

Note that τ_F is a linear and continuous operator from the weak topology of $L^q(\Omega;\mathbf{R}^N)$ to the (unique) topology on F^*. Therefore $G_F: F \to 2^{F^*}$, given by the formula

$$G_F(v_F) := \tau_F \Gamma_R(v_F) \quad \text{for } v_F \in F, \tag{29}$$

is upper semicontinuous.

By the pseudomonotonicity of A it follows that $A_F: F \to F^*$ is continuous. Thus, $A_F + G_F - g_F: F \to 2^{F^*}$ is an upper semicontinuous multivalued mapping with nonempty, bounded, closed and convex values. Moreover, for any $v_F \in F$ and $\psi_F \in G_F(v_F)$ one has

$$\begin{aligned}\langle A_F v_F + \psi_F - g_F, v_F\rangle_F \\ \geq \langle Av_F - g, v_F\rangle_V - \int_\Omega \widetilde{j_R^0}(v_F; -v_F)\, d\Omega.\end{aligned} \tag{30}$$

Hence, in view of Lemma 3, for $R > R_0$ there exists $\mathcal{M} > 0$ not depending on $F \in \Lambda$ such that the condition $\|v_F\|_V = \mathcal{M} + 1$ implies

$$\langle A_F v_F + \psi_F - g_F, v_F\rangle_F \geq 0. \tag{31}$$

Accordingly, one can invoke [1, Corol. 3, p. 337] to deduce the existence of $u_F \in F$ with

$$\|u_F\|_V \leq \mathcal{M} + 1 \tag{32}$$

such that $0 \in A_F u_F + G_F(u_F) - g_F$. This implies that for some $\chi_F \in \Gamma_R(u_F)$ it follows that $\psi_F = \tau_F(\chi_F)$ and (u_F, χ_F) is a solution of (P_F).

In the next step it will be shown that (P_R), $R > R_0$, has solutions.

For $F \in \Lambda$, let

$$\mathcal{W}_F := \bigcup_{\substack{F' \in \Lambda,\\ F' \supset F}} \left\{ u_{F'} \in V : \begin{array}{c}(u_{F'}, \chi_{F'})\\ \text{satisfies } (P_{F'})\\ \text{for some}\\ \chi_{F'} \in L^q(\Omega;\mathbf{R}^N)\end{array} \right\}.$$

The symbol weakcl $(\mathcal{W}_F)$ will be used to denote the closure of $\mathcal{W}_F$ in the weak topology of V. From (32) one gets

$$\text{weakcl}(\mathcal{W}_F) \subset B_V(O, \mathcal{M} + 1), \quad \forall F \in \Lambda,$$

where $B_V(O, \mathcal{M} + 1) := \{v \in V : \|v\|_V \leq \mathcal{M} + 1\}$. Thus, the family $\{\text{weakcl}(\mathcal{W}_F): F \in \Lambda\}$ is contained in the weakly compact set $B_V(O, \mathcal{M} + 1)$ of V. Further, for any $F_1, \ldots, F_k \in \Lambda$, $k = 1, 2, \ldots$, the inclusion $\mathcal{W}_{F_1} \cap \cdots \cap \mathcal{W}_{F_k} \supset \mathcal{W}_F$ results, with $F = F_1 + \cdots + F_k$. Therefore, the family $\{\text{weakcl}(\mathcal{W}_F): F \in \Lambda\}$ has the finite intersection property. This implies that $\cap_{F \in \Lambda}$ weakcl $(\mathcal{W}_F)$ is not empty. From now on, let $u_R \in B_V(0, \mathcal{M} + 1)$ belong to this intersection.

Fix $v \in V$ arbitrarily and choose $F \in \Lambda$ such that u_R, $v \in F$. Thus, there exists a sequence $\{u_{F_n}\} \subset \mathcal{W}_F$ with $u_{F_n} \to u_R$ weakly in V. Let $\chi_{F_n} \in \Gamma_R(u_{F_n})$ denote the corresponding sequence for which (u_{F_n}, χ_{F_n}) is a solution of (P_{F_n}) (for simplicity of notation, the symbols $\{u_n\}$ and $\{\chi_n\}$ will be used instead of u_{F_n} and χ_{F_n}, respectively). Therefore

$$\begin{aligned}\langle Au_n - g, w - u_n\rangle_V + \int_\Omega \chi_n \cdot (w - u_n)\, d\Omega = 0,\\ \forall\, w \in F_n.\end{aligned} \tag{33}$$

Since $\|\chi_n\|_{L^q(\Omega;\mathbf{R}^N)} \leq K_R$ and $L^q(\Omega;\mathbf{R}^N)$ is reflexive, it can also be supposed that for some $\chi_R \in L^q(\Omega;\mathbf{R}^N)$, $\chi_n \to \chi_R$ weakly in $L^q(\Omega;\mathbf{R}^N)$. By the hypothesis, the imbedding $V \subset L^p(\Omega;\mathbf{R}^N)$ is compact, so $u_n \to u_R$ strongly in $L^p(\Omega;\mathbf{R}^N)$. Consequently, by the upper semicontinuity of Γ_R from $L^p(\Omega;\mathbf{R}^N)$ to $L^q(\Omega;\mathbf{R}^N)$ ($L;^q(\Omega;\mathbf{R}^N)$ being endowed with the weak topology) it follows immediately that $\chi_R \in \Gamma_R(u_R)$, i. e. (12) holds. Moreover, $\int_\Omega \chi_n \cdot (u_R - u_n) d\Omega \to 0$ as $n \to \infty$ and (33) with $w = u_R$ lead to $\lim \langle Au_n, u_n - u_R\rangle_V = 0$. Accordingly, the pseudomonotonicity of A allows the conclusion that $\langle Au_n, u_n\rangle_V \to \langle Au_R, u_R\rangle_V$ and $Au_n \to Au_R$

weakly in V^*. Finally, substituting $w = v$ in (33) and letting $n \to \infty$ give in conclusion (11) with $v \in V$ chosen arbitrarily. Thus the existence of solutions of (P_R) has been established.

Let us proceed to the boundedness of solutions $\{u_R\}_{R>R_0}$ of (P_R). Suppose on the contrary that this claim is not true. Then according to (11) and (12) there would exist a sequence $R_n \to \infty$ such that $\| u_{R_n} \|_V \to \infty$ as $n \to \infty$, and

$$\langle Au_{R_n} - g, u_{R_n} \rangle_V - \int_\Omega \widetilde{j^0_{R_n}}(u_{R_n}; -u_{R_n})\, d\Omega \le 0. \quad (34)$$

From now on, for simplicity of notations, instead of the subscript 'R_n' we write 'n'. Eq. (34) allows us to follow the lines of the proof of Lemma 3. First, analogously one arrives at the representation

$$u_n = e_n \left(\frac{1}{e_n} \widehat{u}_n + \theta_n \right),$$

with $\widehat{u}_n / e_n \to 0$ strongly in V and $\theta_n \to \theta_0$ in V_0 as $n \to \infty$ for some $\theta_0 \in V_0$ with $\| \theta_0 \|_V = 1$. Secondly, the counterpart of (21) can be obtained in the form

$$\langle g, \theta \rangle_V \ge \limsup_{n\to\infty} \int_\Omega \left[-\widetilde{j^0_{R_n}} \left(e_n \left(\frac{1}{e_n} \widehat{u}_n + \theta_n \right); -\frac{1}{e_n} \widehat{u}_n - \theta_n \right) \right] d\Omega. \quad (35)$$

But

$$\widetilde{j^0_{R_n}} \left(e_n \left(\frac{1}{e_n} \widehat{u}_n + \theta_n \right); -\frac{1}{e_n} \widehat{u}_n - \theta_n \right) = j^0 \left(e_n \left(\frac{1}{e_n} \widehat{u}_n + \theta_n \right); -\frac{1}{e_n} \widehat{u}_n - \theta_n \right),$$

if $\left| \widehat{u}_n + e_n \theta_n \right| \le R_n$ and

$$\widetilde{j^0_{R_n}} \left(e_n \left(\frac{1}{e_n} \widehat{u}_n + \theta_n \right); -\frac{1}{e_n} \widehat{u}_n - \theta_n \right) = j^0 \left(\frac{R_n}{\left| \frac{1}{e_n} \widehat{u}_n + \theta_n \right|} \left(\frac{1}{e_n} \widehat{u}_n + \theta_n \right); -\frac{1}{e_n} \widehat{u}_n - \theta_n \right),$$

if $\left| \widehat{u}_n + e_n \theta_n \right| > R_n$.. Therefore we easily conclude, using (7), that

$$\limsup_{n\to\infty} -\widetilde{j^0_{R_n}} \left(e_n \left(\frac{1}{e_n} \widehat{u}_n + \theta_n \right); -\frac{1}{e_n} \widehat{u}_n - \theta_n \right) \ge j^\infty(\theta_0).$$

Consequently, by Fatou 's lemma,

$$\langle g, \theta_0 \rangle_V \ge \int_\Omega j^\infty(\theta_0)\, d\Omega,$$

contrary to (6). Thus, the boundedness of $\{u_R\}_{R>R_0}$ follows and the proof of Proposition 4 is complete.

The next result is related to the compactness property of $\{\chi_R : R > R_0\}$ in $L^1(\Omega; \mathbf{R}^N)$.

Proposition 5 *Let a pair $(u_R, \chi_R) \in V \times L^q(\Omega; \mathbf{R}^N)$ be a solution of (P_R). Then the set*

$$\left\{ \chi_R \in L^q(\Omega; \mathbf{R}^N) : \begin{array}{c} (u_R, \chi_R) \\ \text{is a solution of} \\ (P_R) \\ \text{for some } u_R \in V, \\ R > R_0 \end{array} \right\}$$

is weakly precompact in $L^1(\Omega; \mathbf{R}^N)$.

Proof According to the Dunford–Pettis theorem [8] it is sufficient to show that for each $\varepsilon > 0$ a $\delta > 0$ can be determined such that for any $\omega \subset \Omega$ with meas $\omega < \delta$,

$$\int_\omega |\chi_R|\, d\Omega < \varepsilon, \quad R > R_0. \quad (36)$$

Fix $r > 0$ and let $\eta \in \mathbf{R}^N$ be such that $|\eta| \le r$. Then, by (9), from $\chi_R \cdot (\eta - u_R) \le \widetilde{j^0_R}(u_R; \eta - u_R)$ it results that

$$\chi_R \cdot \eta \le \chi_R \cdot u_R + \widetilde{\alpha}(r)(1 + |u_R|^\sigma) \quad (37)$$

a.e. in Ω. Let us set

$$\eta \equiv \frac{r}{\sqrt{N}} (\operatorname{sgn} \chi_{R_1}, \ldots, \operatorname{sgn} \chi_{R_N}),$$

where χ_{R_i}, $i = 1, \ldots, N$, are the components of χ_R and where $\operatorname{sgn} y = 1$ if $y > 0$, $\operatorname{sgn} y = 0$ if $y = 0$, and $\operatorname{sgn} y = -1$ if $y < 0$. It is not difficult to verify that $|\eta| \le r$ for almost all $x \in \Omega$ and that

$$\chi_R \cdot \eta \ge \frac{r}{\sqrt{N}} |\chi_R|.$$

Therefore, by (37) the estimate follows

$$\frac{r}{\sqrt{N}} |\chi_R| \le \chi_R \cdot u_R + \widetilde{\alpha}(r)(1 + |u_R|^\sigma).$$

Integrating this inequality over $\omega \subset \Omega$ yields

$$\int_\omega |\chi_R|\, d\Omega \le \frac{\sqrt{N}}{r}\int_\omega \chi_R \cdot u_R\, d\Omega + \frac{\sqrt{N}}{r}\widetilde{\alpha}(r)\,\mathrm{meas}\,\omega + \frac{\sqrt{N}}{r}\widetilde{\alpha}(r)(\mathrm{meas}\,\omega)^{(p-\sigma)/p}\,\|u_R\|^{\sigma}_{L^p(\Omega)}\, d\Omega. \quad (38)$$

Thus, from (24) one obtains

$$\begin{aligned}\int_\omega |\chi_R|\, d\Omega &\le \frac{\sqrt{N}}{r}\int_\omega \chi_R \cdot u_R\, d\Omega + \frac{\sqrt{N}}{r}\widetilde{\alpha}(r)\,\mathrm{meas}\,\omega \\ &\quad + \frac{\sqrt{N}}{r}\widetilde{\alpha}(r)(\mathrm{meas}\,\omega)^{(p-\sigma)/p}\gamma^{\sigma}\,\|u_R\|^{\sigma}_{V}\, d\Omega \\ &\le \frac{\sqrt{N}}{r}\int_\omega \chi_R \cdot u_R\, d\Omega + \frac{\sqrt{N}}{r}\widetilde{\alpha}(r)\,\mathrm{meas}\,\omega \\ &\quad + \frac{\sqrt{N}}{r}\widetilde{\alpha}(r)(\mathrm{meas}\,\omega)^{(p-\sigma)/p}\gamma^{\sigma} M^{\sigma}\, d\Omega \qquad (39)\end{aligned}$$

($\|\cdot\|_{L^p(\Omega;R^N)} \le \gamma\,\|\cdot\|_V$).

Further, it will be shown that

$$\int_\omega \chi_R \cdot u_R\, d\Omega \le C \qquad (40)$$

for some positive constant C not depending on $\omega \subset \Omega$ and $R > R_0$. Indeed, from (10) one can easily deduce that

$$\chi_R \cdot u_R + k\,|u_R| \ge 0 \quad \text{a.e. in } \Omega.$$

Thus it follows that

$$\int_\omega (\chi_R \cdot u_R + k\,|u_R|)\, d\Omega \le \int_\Omega (\chi_R \cdot u_R + k\,|u_R|)\, d\Omega,$$

and consequently

$$\int_\omega \chi_R \cdot u_R\, d\Omega \le \int_\Omega \chi_R \cdot u_R\, d\Omega + 2k_1\,\|u_R\|_V.$$

But A maps bounded sets into bounded sets. Therefore, by means of (11) and (24),

$$\begin{aligned}\int_\Omega \chi_R \cdot u_R\, d\Omega &= -\langle Au_R - g, u_R\rangle_V \\ &\le \|Au_R - g\|_{V^*}\,\|u_R\|_V \le C_0, \quad C_0 = \text{const},\end{aligned}$$

and consequently, (40) easily follows. Further, from (39) and (40), for $r > 0$,

$$\int_\omega |\chi_R|\, d\Omega \le \frac{\sqrt{N}}{r}C + \frac{\sqrt{N}}{r}\widetilde{\alpha}(r)\,\mathrm{meas}\,\omega + \frac{\sqrt{N}}{r}\widetilde{\alpha}(r)(\mathrm{meas}\,\omega)^{(p-\sigma)/p}\gamma^{\sigma} M^{\sigma}\, d\Omega. \quad (41)$$

This estimate is crucial for obtaining (36). Namely, let $\varepsilon > 0$. Fix $r > 0$ with

$$\frac{\sqrt{N}}{r}C < \frac{\varepsilon}{2} \qquad (42)$$

and determine $\delta > 0$ small enough to fulfill

$$\frac{\sqrt{N}}{r}\widetilde{\alpha}(r)\,\mathrm{meas}\,\omega + \frac{\sqrt{N}}{r}\widetilde{\alpha}(r)(\mathrm{meas}\,\omega)^{(p-\sigma)/p}\gamma^{\sigma} M^{\sigma} \le \frac{\varepsilon}{2},$$

provided that meas $\omega < \delta$. Thus, from (41) it follows that for any $\omega \subset \Omega$ with meas $\omega < \delta$,

$$\int_\omega |\chi_R|\, d\Omega \le \varepsilon, \quad R > R_0. \qquad (43)$$

Finally, $\{\chi_R\}_{R>R_0}$ is equi-integrable and its precompactness in $L^1(\Omega;\mathbf{R}^N)$ has been proved [8].

Now the main result will be formulated.

Theorem 6 *Let $A: V \to V^*$ be a pseudomonotone, bounded operator, $j: \mathbf{R}^N \to \mathbf{R}$ a locally Lipschitz function. Suppose that (1)-(3) and (6) hold. Then there exist $u \in V$ and $\chi \in L^1(\Omega;\mathbf{R}^N)$ such that*

$$\langle Au - g, v - u\rangle_V + \int_\Omega \chi \cdot (v - u)\, d\Omega = 0, \quad \forall v \in V \cap L^\infty(\Omega;\mathbf{R}^N), \quad (44)$$

$$\begin{cases}\chi \in \partial j(u) & \text{a.e. in } \Omega,\\ \chi \cdot u \in L^1(\Omega).\end{cases} \qquad (45)$$

Moreover, the hemivariational inequality holds:

$$\langle Au - g, v - u\rangle_V + \int_\Omega j^0(u; v - u)\, d\Omega \ge 0, \quad \forall v \in V, \quad (46)$$

where the integral above is assumed to take $+\infty$ as its value if $j^0(u;v-u) \notin L^1(\Omega)$.

Proof The proof is divided into a sequence of steps.

Step 1. From Propositions 4 and 5 it follows that from the set $\{ u_R, \chi_R \}_{R>R_0}$ of solutions of (P_R) a sequence $\{ u_{R_n}, \chi_{R_n} \}$ can be extracted with $R_n \to \infty$ as $n \to \infty$ (for simplicity of notations it will be denoted by (u_n, χ_n)), such that

$$\langle Au_n - g, v - u_n \rangle_V + \int_\Omega \chi_n \cdot (v - u_n)\, d\Omega = 0, \forall v \in V, \quad (47)$$

and

$$\begin{cases} \chi_n \in \Gamma_{R_n}(u_n), \\ u_n \to u & \text{weakly in } V, \\ \chi_n \to \chi & \text{weakly in } L^1(\Omega;\mathbf{R}^N) \end{cases} \quad (48)$$

for some $u \in V$ and $\chi \in L^1(\Omega;\mathbf{R}^N)$.

The boundedness of $\{Au_n\}$ in V^* (recall that A has been assumed to be bounded and that $\| u_n \|_V \le M$) allows the conclusion that for some $B \in V^*$,

$$Au_n \to B \quad \text{weakly in } V^* \quad (49)$$

(by passing to a subsequence, if necessary). Thus, (47) implies that the equality

$$\langle B - g, v \rangle_V + \int_\Omega \chi \cdot v\, d\Omega = 0 \quad (50)$$

is valid for any $v \in V \cap L^\infty(\Omega;\mathbf{R}^N)$.

Step 2. Now it will be proved that $\chi \in \partial j(u)$ a.e in Ω, i. e. the first condition in (45) is fulfilled. Since V is compactly imbedded into $L^p(\Omega;\mathbf{R}^N)$, due to (48) one may suppose that

$$u_n \to u \quad \text{strongly in } L^p(\Omega;\mathbf{R}^N). \quad (51)$$

This implies that for a subsequence of $\{u_n\}$ (again denoted by the same symbol) one gets $u_n \to u$ a.e. in Ω. Thus, from Egoroff 's theorem it follows that for any $\varepsilon > 0$ a subset $\omega \subset \Omega$ with meas $\omega < \varepsilon$ can be determined such that $u_n \to u$ uniformly in $\Omega \setminus \omega$ with $u \in L^\infty(\Omega \setminus \omega;\mathbf{R}^N)$. Let $v \in L^\infty(\Omega \setminus \omega;\mathbf{R}^N)$ be an arbitrary function. From the estimate

$$\int_{\Omega\setminus\omega} \chi_n \cdot v\, d\Omega \le \int_{\Omega\setminus\omega} \widetilde{j^0_{R^n}}(u_n; v)\, d\Omega = \int_{\Omega\setminus\omega} j^0(u_n; v)\, d\Omega, \quad \text{(for large } n)$$

(u_n remains pointwise uniformly bounded in $\Omega \setminus \omega$ and $R_n \to \infty$ as $n \to \infty$) combined with the weak convergence in $L^1(\Omega;\mathbf{R}^N)$ of χ_n to χ, (51) and with the upper semicontinuity of

$$L^\infty(\Omega \setminus \omega;\mathbf{R}^N) \ni u_n \longmapsto \int_{\Omega\setminus\omega} j^0(u_n; v)\, d\Omega,$$

it follows that

$$\int_{\Omega\setminus\omega} \chi \cdot v\, d\Omega \le \int_{\Omega\setminus\omega} j^0(u; v)\, d\Omega, \quad \forall v \in L^\infty(\Omega \setminus \omega;\mathbf{R}^N).$$

But the last inequality allows us to state that $\chi \in \partial j(u)$ a.e. in $\Omega \setminus \omega$. Since meas $\omega < \varepsilon$ and ε was chosen arbitrarily,

$$\chi \in \partial j(u) \quad \text{a.e. in } \Omega, \quad (52)$$

as claimed.

Step 3. Now it will be shown that $\chi \cdot u \in L^1(\Omega)$, i. e. the second condition in (45) holds. For this purpose we shall need the following truncation result for vector-valued Sobolev spaces.

Theorem 7 ([20]) *For each $v \in H^1(\Omega;\mathbf{R}^N)$ there exists a sequence of functions $\{ \varepsilon_n \} \subset L^\infty(\Omega)$ with $0 \le \varepsilon_n \le 1$ such that*

$$\begin{aligned} &\{(1-\varepsilon_n)v\} \subset H^1(\Omega;\mathbf{R}^N) \cap L^\infty(\Omega;\mathbf{R}^N) \\ &(1-\varepsilon_n)v \to v \quad \textit{strongly in } H^1(\Omega;\mathbf{R}^N). \end{aligned} \quad (53)$$

Remark 8 For the truncation procedure of the form (53) in the case of a scalar-valued Sobolev space $W^{p,m}(\Omega)$ the reader is referred to [11].

According to the aforementioned theorem, for $u \in V$ one can find a sequence $\{ \varepsilon_k \} \in L^\infty(\Omega)$ with $0 \le \varepsilon_k \le 1$ such that $\widetilde{u}_k := (1-\varepsilon_k)u \in V \cap L^\infty(\Omega;\mathbf{R}^N)$ and $\widetilde{u}_k \to u$ in V as $k \to \infty$. Without loss of generality it can be assumed that $\widetilde{u}_k \to u$ a.e. in Ω. Since it is already

known that $\chi \in \partial j(u)$, one can apply (3) to obtain $\chi \cdot (-u) \le j^0(u;-u) \le k\,|u|$. Hence

$$\chi \cdot \widetilde{u}_k = (1-\varepsilon_k)\chi \cdot u \ge -k\,|u|\,. \tag{54}$$

This implies that the sequence $\{\chi \cdot \widetilde{u}_k\}$ is bounded from below and $\chi \cdot \widetilde{u}_k \to \chi \cdot u$ a.e. in Ω. On the other hand, due to (50) one gets

$$C \ge \langle -B+g, \widetilde{u}_k\rangle_V = \int_\Omega \chi \cdot \widetilde{u}_k \, d\Omega$$

for a positive constant C. Thus, by Fatou 's lemma $\chi \cdot u \in L^1(\Omega)$, as required.

Step 4. Now the inequality

$$\liminf_{n\to\infty} \int_\Omega \chi_n \cdot u_n \, d\Omega \ge \int_\Omega \chi \cdot u \, d\Omega \tag{55}$$

will be established. It can be supposed that $u_n \to u$ a.e. in Ω, because $u_n \to u$ strongly in $L^p(\Omega;\mathbf{R}^N)$. Fix $v \in L^\infty(\Omega;\mathbf{R}^N)$ arbitrarily. Since $\chi_n \in \Gamma_{R_n}(u_n)$, Eq. (9) implies

$$\begin{aligned}\chi_n \cdot (v-u_n) &\le \widetilde{j^0_{R^n}}(u_n; v-u_n)\\ &\le \widetilde{\alpha}(\|v\|_{L^\infty(\Omega;\mathbf{R}^N)})(1+|u_n|^\sigma).\end{aligned} \tag{56}$$

From Egoroff 's theorem it follows that for any $\varepsilon > 0$ a subset $\omega \subset \Omega$ with meas $\omega < \varepsilon$ can be determined such that $u_n \to u$ uniformly in $\Omega \setminus \omega$. One can also suppose that ω is small enough to fulfill $\int_\omega \widetilde{\alpha}(\|v\|_{L^\infty(\Omega;\mathbf{R}^N)})(1+|u_n|^\sigma)\,d\Omega \le \varepsilon$, $n = 1, 2, \ldots$, and $\int_\omega \alpha(\|v\|_{L^\infty(\Omega;\mathbf{R}^N)})(1+\|u^\sigma)\,d\Omega \le \varepsilon$. Hence

$$\begin{aligned}&\int_\Omega \widetilde{j^0_{R^n}}(u_n; v-u_n)\,d\Omega\\ &\le \int_{\Omega\setminus\omega} \widetilde{j^0_{R^n}}(u_n; v-u_n)\,d\Omega + \varepsilon\\ &= \int_{\Omega\setminus\omega} j^0(u_n; v-u_n)\,d\Omega + \varepsilon \quad \text{(for large } n),\end{aligned}$$

which by Fatou 's lemma and upper semicontinuity of $j^0(\cdot;\cdot)$ yields

$$\begin{aligned}&\liminf_{n\to\infty} \int_\Omega -\widetilde{j^0_{R^n}}(u_n; v-u_n)\,d\Omega\\ &\ge \int_\Omega -j^0(u; v-u)\,d\Omega - 2\varepsilon.\end{aligned}$$

By arbitrariness of $\varepsilon > 0$ and (56) one obtains

$$\begin{aligned}&\liminf_{n\to\infty} \int_\Omega \chi_n \cdot u_n \, d\Omega\\ &\ge \int_\Omega \chi \cdot v \, d\Omega - \int_\Omega j^0(u; v-u)\,d\Omega,\\ &\qquad\qquad \forall v \in V \cap L^\infty(\Omega;\mathbf{R}^N).\end{aligned} \tag{57}$$

By substituting $v = \widetilde{u}_k := (1-\varepsilon_k)u$ (with $\widetilde{u}_k$ as described in the truncation argument of Theorem 7) into the right-hand side of (57) one gets

$$\begin{aligned}&\liminf_{n\to\infty} \int_\Omega \chi_n \cdot u_n \, d\Omega\\ &\ge \liminf_{k\to\infty} \int_\Omega \chi \cdot \widetilde{u}_k \, d\Omega\\ &\quad - \limsup_{k\to\infty} \int_\Omega j^0(u; \widetilde{u}_k - u)\,d\Omega.\end{aligned} \tag{58}$$

Taking into account that $\widetilde{u}_k \to u$ a.e. in Ω,

$$j^0(u; \widetilde{u}_k - u) = \varepsilon_k j^0(u;-u) \le \varepsilon_k k\,|u| \le k\,|u|$$

and $|\chi \cdot u| \ge \chi \cdot \widetilde{u}_k = (1-\varepsilon_k)\chi \cdot u \ge -k\,|u|$, Fatou 's lemma and the dominated convergence can be used to deduce

$$\limsup_{k\to\infty} \int_\Omega j^0(u; \widetilde{u}_k - u)\,d\Omega \le 0,$$

and

$$\lim_{k\to\infty} \int_\Omega \chi \cdot \widetilde{u}_k \, d\Omega = \int_\Omega \chi \cdot u \, d\Omega.$$

Finally, combining the last two inequalities with (58) yields (55), as required.

Step 5. The next claim is that

$$\langle B-g, u\rangle_V + \int_\Omega \chi \cdot u \, d\Omega = 0. \tag{59}$$

Indeed, (50) implies

$$\langle B-g, \widetilde{u}_k\rangle_V + \int_\Omega \chi \cdot \widetilde{u}_k \, d\Omega = 0, \tag{60}$$

with $\{\widetilde{u}_k\}$ as in Step 3. Since $\chi \cdot u \in L^1(\Omega)$ and $-k|u| \leq \chi \cdot \widetilde{u}_k = (1-\varepsilon_k)\chi \cdot u \leq |\chi \cdot u|$, by the dominated convergence,

$$\int_\Omega \chi \cdot \widetilde{u}_k \, d\Omega \to \int_\Omega \chi \cdot u \, d\Omega.$$

It means that (59) has to hold by passing to the limit as $k \to \infty$ in (60).

Step 6. In this step it will be shown that the pseudomonotonicity of A and (47) imply (44). Indeed, (47) with $v \in V \cap L^\infty(\Omega;\mathbf{R}^N)$ and (49) allows to state that

$$\limsup_{n\to\infty} \langle Au_n, u_n - u\rangle_V \leq \langle B - g, v - u\rangle_V + \int_\Omega \chi \cdot v \, d\Omega - \liminf_{n\to\infty} \int_\Omega \chi_n \cdot u_n \, d\Omega.$$

Substituting $v = \widetilde{u}_k$ with $\widetilde{u}_k$ as in Step 3 and taking into account (55) one arrives at $\limsup_{n\to\infty} \langle Au_n, u_n - u\rangle_V \leq 0$ (by the application of the limit procedure as $k \to \infty$). Therefore the use of pseudomonotonicity of A is allowed and yields $\langle Au_n, u_n\rangle_V \to \langle Au, u\rangle_V$ and $Au_n \to B = Au$ weakly in V^* as $n \to \infty$. Finally, (47) implies (44), as claimed.

Step 7. In the final step of the proof it will be shown that (44) and (45) imply (46). For this purpose, choose $v \in V \cap L^\infty(\Omega;\mathbf{R}^N)$ arbitrarily. From (2) one has $\chi \cdot (v - u) \leq j^0(u;v-u) \leq \alpha(\|v\|_{L^\infty(\Omega;\mathbf{R}^N)})(1 + |u|^\sigma)$ with $\chi \cdot (v-u) \in L^1(\Omega)$ and $\alpha(\|v\|_{L^\infty(\Omega;\mathbf{R}^N)})(1 + |u|^\sigma) \in L^1(\Omega)$. Hence $j^0(u;v-u)$ is finite integrable and consequently, (46) follows immediately from (44).

Now consider the case $j^0(u;v-u) \in L^1(\Omega)$ with $v \notin V \cap L^\infty(\Omega;\mathbf{R}^N)$. According to Theorem 7 there exists a sequence $\widetilde{v}_k = (1-\varepsilon_k)v$ such that $\{\widetilde{v}_k\} \subset V \cap L^\infty(\Omega;\mathbf{R}^N)$ and $\widetilde{v}_k \to v$ strongly in V. Since

$$\langle Au - g, \widetilde{v}_k - u\rangle_V + \int_\Omega j^0(u;\widetilde{v}_k - u)\, d\Omega \geq 0,$$

so in order to establish (46) it remains to show that

$$\limsup_{k\to\infty} \int_\Omega j^0(u;\widetilde{v}_k - u)\, d\Omega \leq \int_\Omega j^0(u;v-u)\, d\Omega.$$

For this purpose let us observe that $\widetilde{v}_k - u = (1-\varepsilon_k)(v-u) + \varepsilon_k(-u)$ which combined with the convexity of $j^0(u;\cdot)$ yields the estimate

$$j^0(u;\widetilde{v}_k - u) \leq (1-\varepsilon_k)j^0(u;v-u) + \varepsilon_k j^0(u;-u) \leq |j^0(u;v-u)| + k|u|.$$

Thus the application of Fatou 's lemma gives the assertion. Finally, the proof of Theorem 6 is complete.

Remark 9 The analogous result to that of Theorem 6 can be formulated for the hemivariational inequality (46) in which $\int_\Omega(\cdot)\, d\Omega$ is replaced by the boundary integral $\int_\Gamma(\cdot)\, d\Gamma$, provided the imbedding $H^1(\Omega) \subset L^p(\Gamma)$ is compact $(1 < p < (2m-2)/(m-2)$, [12]).

Example 10 Let us consider a linear elastic body which in its undeformed state occupies an open, bounded, connected subset Ω of $\mathbf{R}^3$. Ω is referred to a fixed Cartesian coordinate system $0x_1x_2x_3$ and its boundary Γ is assumed to be Lipschitz regular; $n = (n_i)$ denotes the outward unit normal vector to Γ. We decompose Γ into two disjointed parts Γ_F and Γ_S such that $\Gamma = \overline{\Gamma}_F \cup \overline{\Gamma}_S$. As usual, the symbols u: $\Omega \to \mathbf{R}^3$ and $\sigma : \Omega \to \mathbf{S}^3$ are used to denote the displacement field and the stress tensor field, respectively. Here $\mathbf{S}^3$ stands for the space of all real-valued 3×3 symmetric matrices.

Consider the boundary value problem:

i) The equilibrium equations:

$$\sigma_{ij,j} + b_i = 0 \quad \text{in } \Omega. \tag{61}$$

ii) The displacement—strain relation:

$$\varepsilon_{ij}(u) = \frac{1}{2}(u_{i,j} + u_{j,i}) \quad \text{in } \Omega. \tag{62}$$

iii) *Hook's law*:

$$\sigma_{ij} = C_{ijkl}\varepsilon_{kl}(u) \quad \text{in } \Omega. \tag{63}$$

iv) The *surface traction* conditions

$$\sigma_{ij}n_j = F_i \quad \text{on } \Gamma_F. \tag{64}$$

v) The nonmonotone subdifferential boundary conditions

$$-S \in \partial j(u) \quad \text{on } \Gamma_S. \tag{65}$$

Here, $S = (S_i) = (\sigma_{ij}n_j)$ is the stress vector, and $\partial j(\cdot)$ is the generalized gradient of Clarke of a locally Lipschitz function j: $\mathbf{R}^3 \to \mathbf{R}$; the summation convention over repeated indices holds and the elasticity tensor $C = (C_{ijkl})$ is assumed to satisfy the classical conditions of ellipticity and symmetry [24].

Let $V = H^1(\Omega;\mathbf{R}^3)$. By making use of the standard technique (cf. [24]), Eqs. (61)-(65) lead to the problem

of finding $u \in V$ such as to satisfy the hemivariational inequality

$$\int_{\Omega} C_{ijkl}\varepsilon_{ij}(u)\varepsilon_{kl}(v-u)\,d\Omega - \int_{\Omega} b_i(v_i-u_i)\,d\Omega - \int_{\Gamma_F} F_i(v_i-u_i)\,d\Gamma + \int_{\Gamma_S} j^0(u;v-u)\,d\Gamma \geq 0, \quad \forall v \in V. \quad (66)$$

Define $A\colon V \to V^*$ by

$$\langle Au, v\rangle_V = \int_{\Omega} C_{ijkl}\varepsilon_{ij}(u)\varepsilon_{kl}(v)\,d\Omega, \quad u, v \in V,$$

and let $V_0 := \mathcal{R} = \{\, \rho \in V : \varepsilon_{ij}(\rho) = 0,\ i, j = 1, 2, 3 \,\}$ denote the space of all rigid-body displacements. Then (1) holds (for details see [24, p. 121]). Accordingly, if (2) (with $\sigma < 4$) and (3) are fulfilled and, moreover, the compatibility condition

$$\int_{\Omega} b_i\rho_i\,d\Omega + \int_{\Gamma_F} F_i\rho_i\,d\Gamma < \int_{\Gamma_S} j^{\infty}(\rho)\,d\Gamma$$

is valid for any $\rho \in \mathcal{R} \setminus \{\, 0 \,\}$, then the hypotheses of the theorem mentioned in Remark 9 are satisfied. Consequently, the existence of solutions to the hemivariational inequality (66) is ensured.

References

1. Aubin JP, Ekeland I (1984) Applied nonlinear analysis. Wiley, New York
2. Baiocchi C, Buttazzo G, Gastaldi F, Tomarelli F (1988) General existence theorems for unilateral problems in continuum mechanics. Arch Rational Mechanics Anal 100:149–180
3. Brèzis H (1968) Èquations et inèquations non-lin èaires dans les espaces vèctoriels en dualitè. Ann Inst Fourier Grenoble 18:115–176
4. Brèzis H, Nirenberg L (1978) Characterizations of the ranges of some nonlinear operators and applications to boundary value problems. Ann Scuola Norm Sup Pisa Cl Sci IV V(2):225–326
5. Browder FE, Hess P (1972) Nonlinear mappings of monotone type in Banach spaces. J Funct Anal 11:251–294
6. Clarke FH (1983) Optimization and nonsmooth analysis. Wiley, New York
7. Duvaut G, Lions JL (1972) Les inèquations en mècanique et en physique. Dunod, Paris
8. Ekeland I, Temam R (1976) Convex analysis and variational problems. North-Holland, Amsterdam
9. Fichera G (1972) Boundary value problems in elasticity with unilateral constraints. Handbuch der Physik, vol VIa/2. Springer, Berlin, pp 347–389
10. Goeleven D (1996) Noncoercive variational problems and related topics. Res Notes Math, vol 357. Longman
11. Hedberg LI (1978) Two approximation problems in function space. Ark Mat 16:51–81
12. Kufner A, John O, Fučik S (1977) Function spaces. Academia, Prague
13. Motreanu D, Naniewicz Z (1996) Discontinuous semilinear problems in vector-valued function spaces. Differential Integral Eq 9:581–598
14. Motreanu D, Panagiotopoulos PD (1995) Nonconvex energy functions, related eigenvalue hemivariational inequalities on the sphere and applications. J Global Optim 6:163–177
15. Motreanu D, Panagiotopoulos PD (1996) On the eigenvalue problem for hemivariational inequalities: Existence and multiplicity of solutions. J Math Anal Appl 197:75–89
16. Naniewicz Z (1994) Hemivariational inequalities with functions fulfilling directional growth condition. Appl Anal 55:259–285
17. Naniewicz Z (1994) Hemivariational inequality approach to constrained problems for star-shaped admissible sets. J Optim Th Appl 83:97–112
18. Naniewicz Z (1995) Hemivariational inequalities with functionals which are not locally Lipschitz. Nonlinear Anal 25:1307–1320
19. Naniewicz Z (1995) On variational aspects of some nonconvex nonsmooth global optimization problem. J Global Optim 6:383–400
20. Naniewicz Z (1997) Hemivariational inequalities as necessary conditions for optimality for a class of nonsmooth nonconvex functionals. Nonlinear World 4:117–133
21. Naniewicz Z, Panagiotopoulos PD (1995) Mathematical theory of hemivariational inequalities and applications. M. Dekker, New York
22. Panagiotopoulos PD (1981) Nonconvex superpotentials in the sense of F.H. Clarke and applications. Mechanics Res Comm 8:335–340
23. Panagiotopoulos PD (1983) Noncoercive energy function, hemivariational inequalities and substationarity principles. Acta Mechanics 48:160–183
24. Panagiotopoulos PD (1985) Inequality problems in mechanics and applications. Convex and nonconvex energy functions. Birkhäuser, Basel
25. Panagiotopoulos PD (1991) Coercive and semicoercive hemivariational inequalities. Nonlinear Anal 16:209–231
26. Panagiotopoulos PD (1993) Hemivariational inequalities. Applications in mechanics and engineering. Springer, Berlin
27. Rauch J (1977) Discontinuous semilinear differential equations and multiple valued maps. Proc Amer Math Soc 64:277–282
28. Webb JRL (1980) Boundary value problems for strongly nonlinear elliptic equations. J London Math Soc 21:123–132

Heuristic and Metaheuristic Algorithms for the Traveling Salesman Problem

YANNIS MARINAKIS
Department of Production Engineering and Management, Decision Support Systems Laboratory, Technical University of Crete, Chania, Greece

MSC2000: 90C59

Article Outline

Introduction

The **Traveling Salesman Problem** (**TSP**) is one of the most representative problems in combinatorial optimization. If we consider a salesman who has to visit n cities [46], the Traveling Salesman Problem asks for the shortest tour through all the cities such that no city is visited twice and the salesman returns at the end of the tour back to the starting city. Mathematically, the problem may be stated as follows: Let $G = (V, E)$ be a graph, where V is a set of n nodes and E is a set of arcs, let $C = [c_{ij}]$ be a cost matrix associated with E, where c_{ij} represents the cost of going from city i to city j, $(i, j = 1, \ldots, n)$, the problem is to find a permutation $(i_1, i_2, i_3, \ldots, i_n)$ of the integers from 1 through n that minimizes the quantity $c_{i_1 i_2} + c_{i_2 i_3} + \ldots + c_{i_n i_1}$.

We speak of a *Symmetric TSP*, if for all pairs (i, j), the distance c_{ij} is equal to the distance c_{ji}. Otherwise, we speak of the *Asymmetric TSP* [7]. If the triangle inequality holds ($c_{ij} \leq c_{ii_1} + c_{i_1 j}$, $\forall i, j, i_1$), the problem is said to be metric. If the cities can be represented as points in the plain such that c_{ij} is the Euclidean distance between point i and point j, then the corresponding TSP is called the Euclidean TSP. Euclidean TSP obeys in particular the triangle inequality $c_{ij} \leq c_{ii_1} + c_{i_1 j}$ for all i, j, i_1.

An integer programming formulation of the Traveling Salesman Problem is defined in a complete graph $G = (V, E)$ of n nodes, with node set $V = \{1, \ldots, n\}$, arc set $E = \{(i, j) | i, j = 1, \ldots, n\}$, and nonnegative costs c_{ij} associated with the arcs [8]:

$$c^* = \min \sum_{i \in V} \sum_{j \in V} c_{ij} x_{ij} \tag{1}$$

s.t.

$$\sum_{j \in V} x_{ij} = 1, \quad i \in V \tag{2}$$

$$\sum_{i \in V} x_{ij} = 1, \quad j \in V \tag{3}$$

$$\sum_{i \in S} \sum_{j \in S} x_{ij} \leq |S| - 1, \quad \forall S \subset V, S \neq \emptyset \tag{4}$$

$$x_{ij} \in \{0, 1\}, \quad \text{for all } i, j \in V, \tag{5}$$

where $x_{ij} = 1$ if arc (i,j) is in the solution and 0 otherwise. In this formulation, the objective function clearly describes the cost of the optimal tour. Constraints (2) and (3) are degree constraints: they specify that every node is entered exactly once and left exactly once. Constraints (4) are subtour elimination constraints. These constraints prohibit the formation of subtours, i. e. tours on subsets of less than V nodes. If there was such a subtour on a subset S of nodes, this subtour would contain $|S|$ edges and as many nodes. Constraints (4) would then be violated for this subset since the left-hand side of (4) would be equal to $|S|$ while the right-hand side would be equal to $|S| - 1$. Because of degree constraints, subtours over one node (and hence, over $n - 1$ nodes) cannot occur. For more formulations of the problem see [34,60].

The Traveling Salesman Problem (TSP) is one of the most famous hard combinatorial optimization problems. It has been proven that TSP is a member of the set of NP-complete problems. This is a class of difficult problems whose time complexity is probably exponential. The members of the class are related so that if a polynomial time algorithm was found for one problem, polynomial time algorithms would exist for all of them [41]. However, it is commonly believed that no such polynomial algorithm exists. Therefore, any attempt to construct a general algorithm for finding optimal solutions for the TSP in polynomial time must (probably) fail. That is, for any such algorithm it is possible to construct problem instances for which the execution time grows at least exponentially with the size of the input. Note, however, that time complexity here

refers to the worst case behavior of the algorithm. It can not be excluded that there exist algorithms whose average running time is polynomial. The existence of such algorithms is still an open question. Since 1950s many algorithms have been proposed, developed and tested for the solution of the problem. Algorithms for solving the TSP may be divided into two categories, *exact algorithms* and *heuristic–metaheuristic algorithms*.

Heuristics for the Traveling Salesman Problem

There is a great need for powerful heuristics that find good suboptimal solutions in reasonable amounts of computing time. These algorithms are usually very simple and have short running times. There is a huge number of papers dealing with finding near optimal solutions for the TSP. Our aim is to present the most interesting and efficient algorithms and the most important ones for facing practical problems. In the 1960s, 1970s and 1980s the attempts to solve the Traveling Salesman Problem focused on **tour construction methods** and **tour improvement methods**. In the last fifteen years, metaheuristics, such as **simulated annealing, tabu search, genetic algorithms** and **neural networks**, were introduced. These algorithms have the ability to find their way out of local optima. Heuristics and metaheuristics constitute an increasingly essential component of solution approaches intended to tackle difficult problems, in general, and global and combinatorial problems in particular.

When a heuristic is designed, the question which arises is about the quality of the produced solution. There are three different ways that one may try to answer this question.

1. **Empirical.** The heuristic is applied to a number of test problem instances and the solutions are compared to the optimal values, if there are known, or to lower bounds on these values [33,35].
2. **Worst Case Analysis.** The idea is to derive bounds on the worst possible deviation from the optimum that the heuristic could produce and to devise bad problem instances for which the heuristic actually achieves this deviation [42].
3. **Probabilistic Analysis.** In the probabilistic analysis it is assumed that problem instances are drawn from certain simple probability distributions, and it is, then, proven mathematically that particular solution methods are highly likely to yield near-optimal solutions when the number of cities is large [43].

Tour Construction methods build up a tour step by step. Such heuristics build a solution (tour) from scratch by a growth process (usually a greedy one) that terminates as soon as a feasible solution has been constructed. The problem with construction heuristics is that although they are usually fast, they do not, in general, produce very good solutions. One of the simplest tour construction methods is the **nearest neighborhood** in which, a salesman starts from an arbitrary city and goes to its nearest neighbor. Then, he proceeds from there in the same manner. He visits the nearest unvisited city, until all cities are visited, and then returns to the starting city [65,68].

An extension of the nearest neighborhood method is the **double-side nearest neighborhood method** where the current path can be extended from both of its endnodes. Some authors use the name **Greedy** for Nearest Neighborhood, but it is more appropriately reserved for the special case of the greedy algorithm of matroid theory [39]. Bentley [11] proposed two very fast and efficient algorithms, the **K-d Trees** and the **Lazily Update Priority Queues**. In his paper, it was the first time that somebody suggested the use of data structures for the solution of the TSP. A priority queue contains items with associated values (the priorities) and support operations that [40]:

- remove the highest priority item from the queue and deliver it to the user,
- insert an item,
- delete an item, and
- modify the priority of an item.

The **insertion procedures** [68] take a subtour of V nodes and attempt to determine which node (not already in the subtour) should join the subtour next (the *selection step*) and then determine where in the subtour it should be inserted (the *insertion step*). The most known of these algorithms is the **nearest insertion algorithm**. Similar to the nearest insertion procedure are the **cheapest insertion** [65], the **arbitrary insertion** [12], the **farthest insertion** [65], the **quick insertion** [12], and the **convex hull insertion** [12] algorithms.

There is a number of heuristic algorithms that are designed for speed rather for quality of the tour they construct [40]. The three most known heuristic algo-

rithms of this category are the **Strip algorithm**, proposed by Beardwood et al. [10], the **Spacefilling Curve** proposed by Platzmann and Bartholdi [58] and the **Fast Recursive Partitioning** heuristic proposed by Bentley [11]. The **saving algorithms** are exchange procedures. The most known of them is the **Clarke-Wright algorithm** [17]. Christofides [12,65] suggested a procedure for solving the TSP based on **spanning trees**. He proposed a method of transforming spanning trees to Eulerian graphs.

The **improvement methods** or **local search methods** start with a tour and try to find all tours that are "neighboring" to it and are shorter than the initial tour and, then, to replace it. The tour improvements methods can be divided into three categories according to the type of the neighborhood that they use [64]. Initially, the **constructive neighborhood methods**, which successively add new components to create a new solution, while keeping some components of the current solution fixed. Some of these methods will be presented in the next section where the most known metaheuristics are presented. Secondly, the **transition neighborhood methods**, which are the classic local search algorithms (classic tour improvement methods) and which iteratively move from one solution to another based on the definition of a neighborhood structure. Finally, the **population based neighborhood methods**, which generalize the two previous categories by considering neighborhoods of more than one solution.

The most known of the local search algorithms is the **2-opt heuristic**, in which two edges are deleted and the open ends are connected in a different way in order to obtain a new tour [48]. Note that there is only one way to reconnect the paths. The **3-opt heuristic** is quite similar with the 2-opt but it introduces more flexibility in modifying the current tour, because it uses a larger neighborhood. The tour breaks into three parts instead of only two [48]. In the general case, δ **edges** in a feasible tour are exchanged for δ edges not in that solution as long as the result remains a tour and the length of that tour is less than the length of the previous tour. **Lin-Kernighan method (LK)** was developed by Lin and Kernighan [37,49,54] and for many years was considered to be the best heuristic for the TSP. The **Or-opt procedure**, well known as **node exchange heuristic**, was first introduced by Or [56]. It removes a sequence of up-to-three adjacent nodes and inserts it at another location within the same route. Or-opt can be considered as a special case of 3-opt (three arcs exchanges) where three arcs are removed and substituted by three other arcs. The **GENI algorithm** was presented by Gendreau, Hertz and Laporte [22]. GENI is a hybrid of tour construction and local optimization.

Metaheuristics for the Traveling Salesman Problem

The last fifteen years an incremental amount of metaheuristic algorithms have been proposed. Simulated annealing, genetic algorithms, neural networks, tabu search, ant algorithms, together with a number of hybrid techniques are the main categories of the metaheuristic procedures. These algorithms have the ability to find their way out of local optima. A number of metaheuristic algorithms have been proposed for the solution of the Traveling Salesman Problem. The most important algorithms published for each metaheuristic algorithm are given in the following:

- **Simulated Annealing** (**SA**) belongs [1,2,45,64] to a class of local search algorithms that are known as *threshold accepting algorithms*. These algorithms play a special role within local search for two reasons. First, they appear to be quite successful when applied to a broad range of practical problems. Second, some threshold accepting algorithms such as SA have a stochastic component, which facilitates a theoretical analysis of their asymptotic convergence. Simulated Annealing [3] is a stochastic algorithm that allows random uphill jumps in a controlled fashion in order to provide possible escapes from poor local optima. Gradually the probability allowing the objective function value to increase is lowered until no more transformations are possible. Simulated Annealing owes its name to an analogy with the annealing process in condensed matter physics, where a solid is heated to a maximum temperature at which all particles of the solid randomly arrange themselves in the liquid phase, followed by cooling through careful and slow reduction of the temperature until the liquid is frozen with the particles arranged in a highly structured lattice and minimal system energy. This ground state is reachable only if the maximum temperature is sufficiently high and the cooling sufficiently slow.

Otherwise a meta-stable state is reached. The meta-stable state is also reached with a process known as quenching, in which the temperature is instantaneously lowered. Its predecessor is the so-called Metropolis filter. Simulated Annealing algorithms for the TSP are presented in [15,55,65].

- **Tabu search (TS)** was introduced by Glover [24,25] as a general iterative metaheuristic for solving combinatorial optimization problems. Computational experience has shown that TS is a well established approximation technique, which can compete with almost all known techniques and which, by its flexibility, can beat many classic procedures. It is a form of local neighbor search. Each solution S has an associated set of neighbors $N(S)$. A solution $S' \in N(S)$ can be reached from S by an operation called a *move*. TS can be viewed as an iterative technique which explores a set of problem solutions, by repeatedly making moves from one solution S to another solution S' located in the neighborhood $N(S)$ of S [31]. TS moves from a solution to its best admissible neighbor, even if this causes the objective function to deteriorate. To avoid cycling, solutions that have been recently explored are declared *forbidden or tabu* for a number of iterations. The tabu status of a solution is overridden when certain criteria (*aspiration criteria*) are satisfied. Sometimes, *intensification* and *diversification* strategies are used to improve the search. In the first case, the search is accentuated in the promising regions of the feasible domain. In the second case, an attempt is made to consider solutions in a broad area of the search space. The first Tabu Search algorithm implemented for the TSP appears to be the one described by Glover [23,29]. Limited results for this implementation and variants on it were reported by Glover [26]. Other Tabu Search algorithms for the TSP are presented in [74].
- **Genetic Algorithms (GAs)** are search procedures based on the mechanics of natural selection and natural genetics. The first GA was developed by John H. Holland in the 1960s to allow computers to evolve solutions to difficult search and combinatorial problems, such as function optimization and machine learning [38]. Genetic algorithms offer a particularly attractive approach for problems like traveling salesman problem since they are generally quite effective for rapid global search of large, non-linear and poorly understood spaces. Moreover, genetic algorithms are very effective in solving large-scale problems. Genetic algorithms mimic the evolution process in nature. GAs are based on an imitation of the biological process in which new and better populations among different species are developed during evolution. Thus, unlike most standard heuristics, GAs use information about a population of solutions, called individuals, when they search for better solutions. A GA is a stochastic iterative procedure that maintains the population size constant in each iteration, called a generation. Their basic operation is the mating of two solutions in order to form a new solution. To form a new population, a binary operator called crossover, and a unary operator, called mutation, are applied [61,62]. Crossover takes two individuals, called parents, and produces two new individuals, called offsprings, by swapping parts of the parents. Genetic algorithms for the TSP are presented in [9,51,59,64,67].
- **Greedy Randomized Adaptive Search Procedure - GRASP** [66] is an iterative two phase search method which has gained considerable popularity in combinatorial optimization. Each iteration consists of two phases, a construction phase and a local search procedure. In the construction phase, a randomized greedy function is used to build up an initial solution. This randomized technique provides a feasible solution within each iteration. This solution is then exposed for improvement attempts in the local search phase. The final result is simply the best solution found over all iterations. Greedy Randomized Adaptive Search Procedure algorithms for the TSP are presented in [50,51].
- The use of **Artificial Neural Networks** to find good solutions to combinatorial optimization problems has recently caught some attention. A neural network consists of a network [57] of elementary nodes (neurons) that are linked through weighted connections. The nodes represent computational units, which are capable of performing a simple computation, consisting of a summation of the weighted inputs, followed by the addition of a constant called the threshold or bias, and the application of a non-linear response (activation) function. The result of the computation of a unit constitutes its output. This output is used as an input for the nodes to which

it is linked through an outgoing connection. The overall task of the network is to achieve a certain network configuration, for instance a required input–output relation, by means of the collective computation of the nodes. This process is often called *self–organization*. Neural Networks algorithms for the TSP are presented in [4,6,53,69].

- The **Ant Colony Optimization (ACO)** metaheuristic is a relatively new technique for solving combinatorial optimization problems (COPs). Based strongly on the Ant System (AS) metaheuristic developed by Dorigo, Maniezzo and Colorni [19], ant colony optimization is derived from the foraging behaviour of real ants in nature. The main idea of ACO is to model the problem as the search for a minimum cost path in a graph. Artificial ants walk through this graph, looking for good paths. Each ant has a rather simple behavior so that it will typically only find rather poor-quality paths on its own. Better paths are found as the emergent result of the global cooperation among ants in the colony. An ACO algorithm consists of a number of cycles (iterations) of solution construction. During each iteration a number of ants (which is a parameter) construct complete solutions using heuristic information and the collected experiences of previous groups of ants. These collected experiences are represented by a digital analogue of trail pheromone which is deposited on the constituent elements of a solution. Small quantities are deposited during the construction phase while larger amounts are deposited at the end of each iteration in proportion to solution quality. Pheromone can be deposited on the components and/or the connections used in a solution depending on the problem. Ant Colony Optimization algorithms for the TSP are presented in [16,18,19,70].
- One way to invest extra computation time is to exploit the fact that many local improvement heuristics have random components, even if in their initial tour construction phase. Thus, if one runs the heuristic multiple times he will get different results and can take the best. The **Iterated Lin Kernighan algorithm** (**ILK**) [54] has been proposed by Johnson [39] and it is considered to be one of the best algorithms for obtaining a first local minimum. To improve this local minimum, the algorithm examines other local minimum tours 'near' the current local minimum. To generate these tours, ILK first applies a random and unbiased nonsequential 4-opt exchange to the current local minimum and then optimizes this 4-opt neighbor using the LK algorithm. If the tour obtained by this process is better than the current local minimum then ILK makes this tour the current local minimum and continues from there using the same neighbor generation process. Otherwise, the current local minimum remains as it is and further random 4-opt moves are tried. The algorithm stops when a stopping criterion based either on the number of iterations or the computational time is satisfied. Two other approaches are the **Iterated 3-opt** and the **Chained Lin-Kernighan** [5], where random kicks are generated from the solution and from these new points the exploration for a better solution is continued [40].
- **Ejection Chain Method** provides a wide variety of reference structures, which have the ability to generate moves not available to neighborhood search approaches traditionally applied to TSP [63,64]. Ejection Chains are variable depth methods that generate a sequence of interrelated simple moves to create a more complex compound move. An ejection consists of a succession of operations performed on a given set of elements, where the m_t operation changes the state of one or more elements which are said to be ejected in the m_t+1 operations. Of course, there is a possibility to appear changes in the state of other elements, which will lead to other ejections, until no more operations can be made [27]. Other Ejection Chain Algorithms are presented in [20,21].
- **Scatter Search** is an evolutionary strategy originally proposed by Glover [28,30]. Scatter Search operates on a set of reference solutions to generate a new set of solutions by weighted linear combinations of structured subset of solutions. The reference set is required to be made up of high quality and diverse solutions and the goal is to produce weighted centers of selected subregions that project these centers into regions of the solution space that are to be explored by auxiliary heuristic procedures.
- **Path Relinking** [28,30], combines solutions by generating paths between them using local search neighborhoods, and selecting new solutions encountered along these paths.

- **Guided Local Search** (**GLS**), originally proposed by Voudouris and Chang [71,72], is a general optimization technique suitable for a wide range of combinatorial optimization problems. The main focus is on the exploitation of problem and search–related information to effectively guide local search heuristics in the vast search spaces of NP-hard optimization problems. This is achieved by augmenting the objective function of the problem to be minimized with a set of penalty terms which are dynamically manipulated during the search process to steer the heuristic to be guided. GLS augments the cost function of the problem to include a set of penalty terms and passes this, instead of the original one, for minimization by the local search procedure. Local search is confined by the penalty terms and focuses attention on promising regions of the search space. Iterative calls are made to local search. Each time local search gets caught in a local minimum, the penalties are modified and local search is called again to minimize the modification cost function. Guided Local Search algorithms for the TSP are presented in [71,72].
- **Noising Method** was proposed by Charon and Hudry [13] and is a metaheuristic where if it is wanted to minimize the function f^1, this method do not take the true values of f^1 into account but it considers that they are perturbed in some way by noises in order to get a noised function ${f^1}_{\text{noised}}$. During the run of the algorithm, the range of the perturbing noises decreases (typically to zero), so that, at the end, there is no significant noise and the optimization of ${f^1}_{\text{noised}}$ leads to the same solution as the one provided by a descent algorithm applied to f^1 with the same initial solution. This algorithm was applied to the Traveling Salesman Problem by Charon and Hudry [14].
- **Particle Swarm Optimization (PSO)** is a population-based swarm intelligence algorithm. It was originally proposed by Kennedy and Eberhart as a simulation of the social behavior of social organisms such as bird flocking and fish schooling [44]. PSO uses the physical movements of the individuals in the swarm and has a flexible and well-balanced mechanism to enhance and adapt to the global and local exploration abilities. PSO algorithms for the solution of the Traveling Salesman Problem are presented in [32,47,73].
- **Variable Neighborhood Search (VNS)** is a metaheuristic for solving combinatorial optimization problems whose basic idea is systematic change of neighborhood within a local search [36]. Variable Neighborhood Search algorithms for the TSP are presented in [52].

References

1. Aarts E, Korst J (1989) Simulated Annealing and Boltzmann Machines - A stochastic Approach to Combinatorial Optimization and Neural Computing. John Wiley and Sons, Chichester
2. Aarts E, Ten Eikelder HMM (2002) Simulated Annealing. In: Pardalos PM, Resende MGC (eds) Handbook of Applied Optimization. Oxford University Press, Oxford, pp 209–221
3. Aarts E, Korst J, Van Laarhoven P (1997) Simulated Annealing. In: Aarts E, Lenstra JK (eds) Local Search in Combinatorial Optimization. John Wiley and Sons, Chichester, pp 91–120
4. Ansari N, Hou E (1997) Computational Intelligence for Optimization, 1st edn. Kluwer, Boston
5. Applegate D, Cook W, Rohe A (2003) Chained Lin-Kernighan for Large Traveling Salesman Problems. Informs J Comput 15:82–92
6. Bai Y, Zhang W, Jin Z (2006) An New Self-Organizing Maps Strategy for Solving the Traveling Salesman Problem. Chaos Solitons Fractals 28(4):1082–1089
7. Balas E, Fischetti M (2002) Polyhedral Theory for the Assymetric Traveling Salesman Problem. In: Gutin G, Punnen A (eds) The Traveling Salesman Problem and its Variations. Kluwer, Dordrecht, pp 117–168
8. Balas E, Toth P (1985) Branch and Bound Methods. In: Lawer EL, Lenstra JK, Rinnoy Kan AHG, Shmoys DB (eds) The Travelling Salesman Problem: A Guided Tour of Combinatorial Optimization. John Wiley and Sons, Chichester, pp 361–401
9. Baralia R, Hildago JI, Perego R (2001) A Hybrid Heuristic for the Traveling Salesman Problem. IEEE Trans Evol Comput 5(6):1–41
10. Beardwood J, Halton JH, Hammersley JM (1959) The Shortest Path Through Many Points. Proc Cambridge Philos Soc 55:299–327
11. Bentley JL (1992) Fast Algorithms for Geometric Traveling Salesman Problems. ORSA J Comput 4:387–411
12. Bodin L, Golden B, Assad A, Ball M (1983) The State of the Art in the Routing and Scheduling of Vehicles and Crews. Comput Oper Res 10:63–212
13. Charon I, Hudry O (1993) The Noising Method: A New Combinatorial Optimization Method. Oper Res Lett 14:133–137
14. Charon I, Hudry O (2000) Applications of the Noising Method to the Traveling Salesman Problem. Eur J Oper Res 125:266–277

15. Chen Y, and Zhang P (2006) Optimized Annealing of Traveling Salesman Problem from the nth-Nearest-Neighbor Distribution. Physica A: Stat Theor Phys 371(2):627–632
16. Chu SC, Roddick JF, Pan JS (2004) Ant Colony System with Communication Strategies. Inf Sci 167(1–4):63–76
17. Clarke G, and Wright J (1964) Scheduling of Vehicles from a Central Depot to a Number of Delivery Points. Oper Res 12:568–581
18. Dorigo M, Gambardella LM (1997) Ant Colony System: A Cooperative Learning Approach to the Traveling Salesman Problem. IEEE Trans Evol Comput 1(1):53–66
19. Dorigo M, Stutzle T (2004) Ant Colony Optimization, A Bradford Book. The MIT Press Cambridge, Massachusetts, London
20. Gamboa D, Rego C, Glover F (2005) Data Structures and Ejection Chains for Solving Large-Scale Traveling Salesman Problems. Eur J Oper Res 160(1):154–171
21. Gamboa D, Rego C, Glover F (2006) Implementation Analysis of Efficient Heuristic Algorithms for the Traveling Salesman Problem. Comput Oper Res 33(4):1154–1172
22. Gendreau M, Hertz A, Laporte G (1992) New Insertion and Postoptimization Procedures for the Traveling Salesman Problem. Oper Res 40:1086–1094
23. Glover F (1986) Future Paths for Integer Programming and Links to Artificial Intelligence. Comput Oper Res 13:533–549
24. Glover F (1989) Tabu Search I. ORSA J Comput 1(3):190–206
25. Glover F (1990) Tabu Search II. ORSA J Comput 2(1):4–32
26. Glover F (1990) Tabu search: A tutorial. Center for Applied Artificial Intelligence, University of Colorado, pp 1–47
27. Glover F (1992) Ejection Chains, Reference Structures and Alternating Path Algorithms for Traveling Salesman Problem. Discrete Appl Math 65:223–253
28. Glover F (1997) A Template for Scatter Search and Path Relinking. Lecture Notes in Computer Science, vol 1363. pp 13–54
29. Glover F, and Laguna M (2002) Tabu Search. In: Pardalos PM, Resende MGC (eds) Handbook of Applied Optimization. Oxford University Press, Oxford, pp 194–209
30. Glover F, Laguna M, Marti R (2003) Scatter Search and Path Relinking: Advances and Applications. In: Glover F, Kochenberger GA (eds) Handbook of Metaheuristics. Kluwer, Boston, pp 1–36
31. Glover F, Laguna M, Taillard E, de Werra D (eds) (1993) Tabu Search. J.C. Baltzer AG, Science Publishers, Basel, Switzerland
32. Goldbarg EFG, Souza GR, Goldbarg MC (2006) Particle Swarm Optimization for the Traveling Salesman Problem. EVOCOP 2006 LNCS 3906:99–110
33. Golden BL, Stewart WR (1985) Empirical Analysis of Heuristics. In: Lawer EL, Lenstra JK, Rinnoy Kan AHG, Shmoys DB (eds) The Traveling Salesman Problem: A Guided Tour of Combinatorial Optimization. John Wiley and Sons, Chichester, pp 207–249
34. Gutin G, Punnen A (eds) (2002) The Traveling Salesman Problem and its Variations. Kluwer, Dordrecht
35. Haimovich M, Rinnoy Kan AHG, Stougie L (1988) Analysis of Heuristics for Vehicle Routing Problems. In: Golden BL, Assad AA (eds) Vehicle Routing: Methods and Studies. Elsevier Science Publishers, North Holland, pp 47–61
36. Hansen P, Mladenovic N (2001) Variable Neighborhood Search: Principles and Applications. Eur J Oper Res 130:449–467
37. Helsgaun K (2000) An Effective Implementation of the Lin-Kernighan Traveling Salesman Heuristic. Eur J Oper Res 126:106–130
38. Holland J H. (1975) Adaptation in Natural and Artificial Systems. University of Michigan Press, Ann Arbor
39. Johnson DS, McGeoch LA (1997) The Traveling Salesman Problem: A Case Study. In: Aarts E, Lenstra JK (eds) Local Search in Combinatorial Optimization. John Wiley and Sons, Chichester, pp 215–310
40. Johnson DS, McGeoch LA (2002) Experimental Analysis of the STSP. In: Gutin G, Punnen A (eds) The Traveling Salesman Problem and its Variations. Kluwer, Dordrecht, pp 369–444
41. Johnson DS, Papadimitriou CH (1985) Computational Complexity. In: Lawer EL, Lenstra JK, Rinnoy Kan AHD, Shmoys DB (eds) The Traveling Salesman Problem: A Guided Tour of Combinatorial Optimization. John Wiley and Sons, Chichester, pp 37–85
42. Johnson DS, Papadimitriou CH (1985) Performance Guarantees for Heuristics. In: Lawer EL, Lenstra JK, Rinnoy Kan AHD, Shmoys DB (eds) The Traveling Salesman Problem: A Guided Tour of Combinatorial Optimization. John Wiley and Sons, Chichester, pp 145–181
43. Karp RM, Steele JM (1985) Probabilistic Analysis of Heuristics. In: Lawer EL, Lenstra JK, Rinnoy Kan AHD, Shmoys DB (eds) The Traveling Salesman Problem: A Guided Tour of Combinatorial Optimization. John Wiley and Sons, Chichester, pp 181–206
44. Kennedy J, Eberhart R (1995) Particle Swarm Optimization. Proc. 1995 IEEE Int Conf Neural Netw 4:1942–1948
45. Kirkpatrick S, Gelatt CD, Vecchi MP (1982) Optimization by Simulated Annealing. Science 220:671–680
46. Lawer EL, Lenstra JK, Rinnoy Kan AHG, Shmoys DB (1985) The Traveling Salesman Problem: A Guided Tour of Combinatorial Optimization. Wiley and Sons, New York
47. Li X, Tian P, Hua J, Zhong N (2006) A Hybrid Discrete Particle Swarm Optimization for the Traveling Salesman Problem. SEAL 2006, LNCS 4247:181–188
48. Lin S (1965) Computer Solutions of the Traveling Salesman Problem. Bell Syst Tech J 44:2245–2269
49. Lin S, Kernighan BW (1973) An Effective Heuristic Algorithm for the Traveling Salesman Problem. Oper Res 21:498–516
50. Marinakis Y, Migdalas A, Pardalos PM (2005) Expanding Neighborhood GRASP for the Traveling Salesman Problem. Comput Optim Appl 32:231–257

51. Marinakis Y, Migdalas A, Pardalos PM (2005) A Hybrid Genetic-GRASP algortihm Using Langrangean Relaxation for the Traveling Salesman Problem. J Combinat Optim 10:311–326
52. Mladenovic N, Hansen P (1997) Variable Neighborhood Search. Comput Oper Res 24:1097–1100
53. Modares A, Somhom S, Enkawa T (1999) A Self-Organizing Neural Network Approach for Multiple Traveling Salesman and Vehicle Routing Problems. Int Trans Oper Res 6(6):591–606
54. Neto DM (1999) Efficient Cluster Compensation for Lin - Kernighan Heuristics. Ph.D. Thesis, Computer Science University of Toronto, Canada
55. Ninio M, Schneider JJ (2005) Weight Annealing. Physica A: Stat Theor Phys 349(3–4):649–666
56. Or I (1976) Traveling Salesman-Type Combinatorial Problems and their Relation to the Logistics of Regional Blood Banking. Ph.D. Thesis, Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston IL
57. Sodererg B, Peterson C (1997) Artificial Neural Networks. In: Aarts E, Lenstra JK (eds) Local Search in Combinatorial Optimization. John Wiley and Sons, Chichester, pp 173–214
58. Platzmann LK, Bartholdi JJ (1989) Spacefilling Curves and the Planar Traveling Salesman Problem. J Assoc Comput Mach 36:719–735
59. Potvin J Y. (1996) Genetic Algorithms for the Traveling Salesman Problem. Metaheuristics Combinatorial Optim Ann Oper Res 63:339–370
60. Punnen AP (2002) The Traveling Salesman Problem: Applications, Formulations and Variations. In: Gutin G, Punnen A (eds) The Traveling Salesman Problem and its Variations. Kluwer, Dordrecht, pp 1–28
61. Reeves CR (1995) Genetic Algorithms. In: Reeves CR (ed) Modern Heuristic Techniques for Combinatorial Problems. McGraw - Hill, London, pp 151–196
62. Reeves CR (2003) Genetic Algorithms. In: Glover F, Kochenberger GA (eds) Handbooks of Metaheuristics. Kluwer, Dordrecht, pp 55–82
63. Rego C (1998) Relaxed Tours and Path Ejections for the Traveling Salesman Problem. Eur J Oper Res 106:522–538
64. Rego C, Glover F (2002) Local Search and Metaheuristics. In: Gutin G, Punnen A (eds) The Traveling Salesman Problem and its Variations. Kluwer, Dordrecht, pp 309–367
65. Reinelt G (1994) The Traveling Salesman, Computational Solutions for TSP Applications. Springer, Berlin
66. Resende MGC, Ribeiro CC (2003) Greedy Randomized Adaptive Search Procedures. In: Glover F, Kochenberger GA (eds) Handbook of Metaheuristics. Kluwer, Boston, pp 219–249
67. Ronald S (1995) Routing and Scheduling Problems. In: Chambers L (ed) Practical Handbook of Genetic Algorithms. CRC Press, New York, pp 367–430
68. Rosenkratz DJ, Stearns RE, Lewis PM (1977) An Analysis of Several Heuristics for the Travelling Salesman Problem. SIAM J Comput 6:563–581
69. Siqueira PH, Teresinha M, Steiner A, Scheer S (2007) A New Approach to Solve the Traveling Salesman Problem. Neurocomputing 70(4–6):1013–1021
70. Taillard ED (2002) Ant Systems. In: Pardalos PM, Resende MGC (eds) Handbook of Applied Optimization. Oxford University Press, Oxford, pp 130–138
71. Voudouris C, Tsang E (1999) Guided Local Search and its Application to the Travelling Salesman Problem. Eur J Oper Res 113:469–499
72. Voudouris C, Tsang E (2003) Guided Local Search. In: Glover F, Kochenberger GA (eds) Handbooks of Metaheuristics. Kluwer, Dordrecht, pp 185–218
73. Wang Y, Feng XY, Huang YX, Pu DB, Zhou WG, Liang YC, Zhou CG (2007) A Novel Quantum Swarm Evolutionary Algorithm and its Applications. Neurocomputing 70 (4–6):633–640
74. Zachariasen M, Dam M (1996) Tabu Search on the Geometric Traveling Salesman Problem. In: Osman IH, Kelly JP (eds) Meta-heuristics: Theory and Applications. Kluwer, Boston, pp 571–587

Heuristic Search

ALEXANDER REINEFELD
ZIB Berlin, Berlin, Germany

MSC2000: 68T20, 90B40, 90C47

Article Outline

Keywords and Phrases

Optimization; Heuristic search

Introduction

Heuristic search [7,9] is a common technique for finding a solution in a decision tree or graph containing one

or more solutions. Many applications in operations research and artificial intelligence rely on heuristic search as their primary solution method.

Heuristic search techniques can be classified into two broad categories: depth-first search (DFS) and best-first search (BFS). As a consequence of its better information base, BFS usually examines fewer nodes but occupies more storage space for maintaining the already explored nodes.

Depth-First Search

DFS expands an initial state by generating its immediate successors. At each subsequent step, one of the most recently generated successors is selected and expanded. At terminal states, or when it can be determined that the current state does not lead to a solution, the search backtracks, that is, the node expansion proceeds with the next most recently generated state. Practical implementations use a stack data structure for maintaining the states (nodes) on the path to the currently explored state. The space complexity of the stack, $O(d)$, increases only linearly with the search depth d.

Backtracking is the most rudimentary variant of DFS. It terminates as soon as any solution has been found; hence, there is no guarantee for finding an optimal (least-cost) solution. Moreover, backtracking might not terminate in graphs containing cycles or when the search depth is unbounded.

Depth-first branch and bound (DFBB) [6] employs a heuristic function to eliminate parts of the search space that cannot contain an optimal solution. It continues after finding a first solution until the search space is completely exhausted. Whenever a better solution is found, the current solution path and its value are updated. Inferior subtrees, i. e., subtrees that are known to be worse than the current solution, are eliminated.

The alpha-beta algorithm [2] used in game tree searching is a variant of DFBB that operates on trees with alternating levels of AND and OR nodes [5]. Because the strength of play correlates to the depth of the search, much effort has been spent on devising efficient parallel implementations (► parallel heuristic search).

Best-First Search

BFS sorts the sequence of node expansions according to a heuristic function. The A* search algorithm [7] uses a heuristic evaluation function $f(n) = g(n) + h(n)$ to decide which successor node n to expand next. Here, $g(n)$ is the cost of the path from the initial state to the current node n and $h(n)$ is the estimated completion cost to a nearest goal state. If h does not overestimate the remaining cost, A* is guaranteed to find an optimal (least-cost) solution: it is said to be admissible. It does so with a minimal number of node expansions [9]—no other search algorithm (with the same heuristic h) can do better. This is possible, because A* keeps the search graph in memory, occupying $O(w^d)$ memory cells for trees of width w and depth d.

Best-first frontier search [4] also finds an optimal solution, but with a much lower space complexity than A*. It only keeps the frontier nodes in memory and discards the interior (closed) nodes. Care must be taken to ensure that the search frontier does not contain gaps that would allow the search to leak back into interior regions. The memory savings are most pronounced in directed acyclic graphs. In the worst case, that is, in trees of width w, it still saves a fraction of 1/w of the nodes that BFS would need to store.

Iterative-deepening A* (IDA*) [3] simulates A*'s best-first node expansion by a series of DFSs, each with the cost-bound $f(n)$ increased by the minimal amount. The cost-bound is initially set to the heuristic estimate of the root node, h (root). Then, for each iteration, the bound is increased to the minimum value that exceeded the previous bound. Like A* , IDA* is guaranteed to find an optimal solution [3], provided the heuristic estimate function h is admissible and never overestimates the path to the goal. IDA* obeys the same asymptotic branching factor as A* [7], if the number of newly expanded nodes grows exponentially with the search depth [3]. This growth rate, the heuristic branching factor, depends on the average number of applicable operators per node and the discrimination power of the heuristic function h.

Applications

Typical applications of heuristic search techniques may be found in many areas—not only in the fields of artificial intelligence and operations research, but also in other parts of computer science.

In the two-dimensional rectangular cutting-stock problem [1], we are given a set $R_s = \{(l_i, w_i), i = 1,\ldots, m\}$

of m rectangles of width w_i and length l_i that are to be cut out of a single rectangular stock sheet S. Assuming that S is of width W and that the theoretically unbounded length is L, the problem is to find an optimal cut with minimal length expansion. Since the elements R_i are cut after the cutting pattern has been determined, we can look at the problem as a bin-packing or vehicle-routing problem, which are also known to be nondeterministic polynomial-time (NP) complete [8].

Very large scale integration (VLSI) floorplan optimization is a stage in the design of VLSI chips, where the dimensions of the basic building blocks (cells) must be determined, subject to the minimization of the total chip layout area. This can be done with a BFS or a DFBB approach. Again, only small problem cases can be solved optimally, because VLSI floorplan optimization is also NP-complete.

In the satisfiability problem, it must be determined whether a Boolean formula containing binary variables in conjunctive normal form is satisfiable, that is, whether an assignment of truth values to the variables exists for which the formula is true.

The 15-puzzle benchmark in single-agent game-tree search consists of 15 square tiles located in a square tray of size 4×4. One square, the "blank square," is kept empty so that an orthogonally adjacent tile can slide into its position, thus leaving an empty position at its origin. The problem is to re-arrange a given initial configuration with the fewest number of moves into a goal configuration without lifting one tile over another. While it would seem easy to obtain any solution, finding optimal (shortest) solutions is NP-complete. The 15-puzzle spawns a search space of $16! \approx 2{\cdot}10^{13}$ states.

See also

- ► Asynchronous Distributed Optimization Algorithms
- ► Automatic Differentiation: Parallel Computation
- ► Load Balancing for Parallel Optimization Techniques
- ► Parallel Computing: Complexity Classes
- ► Parallel Computing: Models
- ► Parallel Heuristic Search
- ► Stochastic Network Problems: Massively Parallel Solution

References

1. Christofides N, Whitlock C (1977) An algorithm for two-dimensional cutting problems. Oper Res 25(1):30–44
2. Knuth DE, Moore RW (1975) An analysis of alpha-beta pruning. Artif Intell 6(4):293–326
3. Korf RE (1985) Depth-first iterative-deepening: An optimal admissible tree search. Artif Intell 27:97–109
4. Korf RE, Zhang W, Thayer I, Hohwald H (2005) Frontier Search J ACM 52:715–748
5. Kumar V, Nau DS, Kanal L (1988) A general branch-and-bound formulation for AND/OR graph and game-tree search. In: Kanal L, Kumar V (eds) Search in Artificial Intelligence. Springer, New York, pp 91–130
6. Lawler EL, Wood DE (1966) Branch and bound methods: A survey. Oper Res 14:600–719
7. Nilsson NJ (1980) Principles of artificial intelligence. Tioga Publ., Palo Alto
8. Papadimitriou CH, Steiglitz K (1982) Combinatorial optimization: Algorithms and complexity. Prentice-Hall, Englewood Cliffs, NJ
9. Pearl J (1984) Heuristics. Intelligent search strategies for computer problem solving. Addison-Wesley, Reading

Heuristics for Maximum Clique and Independent Set

MARCELLO PELILLO
University Ca' Foscari di Venezia,
Venezia Mestre, Italy

MSC2000: 90C59, 05C69, 05C85, 68W01

Article Outline

Keywords

Heuristics; Algorithms; Clique; Independent set

Throughout this article, $G = (V, E)$ is an arbitrary undirected and weighted graph unless otherwise specified, where $V = (1, \ldots, n\}$ is the vertex set of G and $E \subseteq V \times V$ is its edge set. For each vertex $i \in V$, a positive weight w_i is associated with i, collected in the *weight vector* $w \in \mathbf{R}^n$. For a subset $S \subseteq V$, the weight of S is defined as $W(S) = \sum_{i \in S} w_i$, and $G(S) = (S, E \cap S \times S)$ is the *subgraph induced by S*. The *cardinality* of S, i. e., the number of its vertices, will be denoted by $|S|$.

A graph $G = (V, E)$ is *complete* if all its vertices are pairwise adjacent, i. e. $\forall i, j \in V$ with i $\neq j$, we have $(i, j) \in E$. A *clique* C is a subset of V such that $G(C)$ is complete. The *clique number* of G, denoted by $\omega(G)$ is the cardinality of the maximum clique. The maximum clique problem asks for cliques of maximum cardinality. The maximum weight clique problem asks for cliques of maximum weight. Given the weight vector $w \in \mathbf{R}^n$, the *weighted clique number* is the total weight of the maximum weight clique, and will be denoted by $\omega(G, w)$.

We should distinguish a *maximum clique* from a *maximal clique*. A maximal clique is one that is not a proper subset of any other clique. A maximum (weight) clique is a maximal clique that has the maximum cardinality (weight).

An *independent set* (also called *stable set* or *vertex packing*) is a subset of V whose elements are pairwise nonadjacent. The maximum independent set problem asks for an independent set of maximum cardinality. The size of a maximum independent set is the *stability number* of G, (denoted by $\alpha(G)$). The maximum weight independent set problem asks for an independent set of maximum weight. Given the weight vector w $\in \mathbf{R}^n$, the *weighted stability number*, denoted $\alpha(G, w)$, is the weight of the maximum weight independent set.

The *complement graph* of $G = (V, E)$ is the graph $\overline{G} = (V, \overline{E})$, where $\overline{E} = \{(i, j) : i, j \in V, i \neq j \text{ and } (i, j) \notin E\}$. It is easy to see that S is a clique of G if and only if S is an independent set of $\overline{G}$. Any result or algorithm obtained for one of the two problems has its equivalent forms for the other one. Hence $\alpha(G) = \omega(\overline{G})$, more generally, $\alpha(G, w) = \omega(\overline{G}, w)$.

The maximum clique and independent set problems are well-known examples of intractable combinatorial optimization problems [18]. Apart from the theoretical interest around these problems, they also find practical applications in such diverse domains as computer vision, experimental design, information retrieval, fault tolerance, etc. Moreover, many important problems turn out to be easily reducible to them, and these include, for example, the Boolean satisfiability problem, the subgraph isomorphism problem, and the vertex covering problem. The maximum clique problem has also a certain historical value, as it was one of the first problems shown to be *NP*-complete in the now classical paper of R.M. Karp on computational complexity [64].

Due to their inherent computational complexity, exact algorithms are guaranteed to return a solution only in a time which increases exponentially with the number of vertices in the graph, and this makes them inapplicable even to moderately large problem instances. Moreover, a series of recent theoretical results show that the problems are in fact difficult to solve even in terms of approximation. Strong evidence of this fact came in 1991, when it was proved in [32] that if there is a polynomial time algorithm that approximates the maximum clique within a factor of $2^{log^{1-\epsilon} n}$, then any *NP*-hard problem can be solved in 'quasipolynomial' time (i. e., in $2^{log^{O(1)} n}$ time). The result was further refined in [6,7] one year later. Specifically, it was proved that there exists an $\epsilon > 0$ such that no polynomial time algorithm can approximate the size of the maximum clique within a factor of n^{ϵ}, unless $P = NP$. Developments along these lines can be found in [14,15,49].

In light of these negative results, much effort has recently been directed towards devising efficient heuristics for maximum clique and independent set, for which no formal guarantee of performance may be provided, but are anyway of interest in practical application. Lacking (almost by definition) a general theory of how these algorithms work, their evaluation is essentially based on massive experimentation. In order to facilitate comparisons among different heuristics, a set of benchmark graphs arising from different applications and problems has recently been constructed in conjunction with the 1993 DIMACS challenge on cliques, coloring and satisfiability [63].

In this article we provide an informal survey of recent heuristics for maximum clique and related problems, and up-to-date bibliographic pointers to the relevant literature. A more comprehensive review and bibliography can be found in [18].

Sequential Greedy Heuristics

Many approximation algorithms in the literature for the maximum clique problem are called *sequential greedy heuristics*. These heuristics generate a maximal clique through the repeated addition of a vertex into a partial clique, or the repeated deletion of a vertex from a set that is not a clique. Decisions on which vertex to be added in or moved out next are based on certain indicators associated with candidate vertices as, for example, the vertex degree. There is also a distinction between heuristics that update the indicators every time a vertex is added in or moved out, and those that do not. Examples of such heuristics can be found in [62,89]. The differences among these heuristics are their choice of indicators and how indicators are updated. A heuristic of this type can run very fast.

Local Search Heuristics

Let us define C_G to be the set of all the maximal cliques of G. Basically, a sequential greedy heuristic finds one set in C_G, hoping it is (close to) the optimal set, and stops. A possible way to improve our approximation solutions is to expand the search in C_G. For example, once we find a set $S \in C_G$, we can search its 'neighbors' to improve S. This leads to the class of the *local search heuristics* [2]. Depending on the neighborhood structure and how the search is performed, different local search heuristics result.

A well-known class of local search heuristics in the literature is the *k-interchange heuristics*. They are based on the *k-neighbor* of a feasible solution. In the case of the maximum clique problem, a set $C \in C_G$ is a k-neighbor of S if $|C \Delta S| \leq k$, where $k \leq |S|$. A k-interchange heuristic first finds a maximal clique $S \in C_G$, then it searches all the k-neighbors of S and returns the best clique found. Clearly, the main factors for the complexity of this class of heuristics are the size of the neighborhood and the searches involved. For example, in the k-interchange heuristic, the complexity grows roughly with $O(n^k)$.

A class of heuristics designed to search various sets of C_G is called the *randomized heuristics*. The main ingredient of this class of heuristics is the part that finds a random set in C_G. A possible way to do that is to include some random factors in the generation of a set of C_G. A randomized heuristic runs a heuristic (with random factors included) a number of times to find different sets over C_G. For example, we can randomize a sequential greedy heuristic and let it run N times. The complexity of a randomized heuristic depends on the complexity of the heuristic and the number N.

An elaborated implementation of the randomized heuristic for the maximum independent set problem can be found in [33], where local search is combined with randomized heuristic. The computational results in it indicated that the approach was effective in finding large cliques of randomly generated graphs. A different implementation of a randomized algorithm for the maximum independent set problem can be found in [5].

Advanced Search Heuristics

Local search algorithms are only capable of finding *local solutions* of an optimization problem. Powerful variations of the basic local search procedure have been developed which try to avoid this problem, many of which are inspired from various phenomena occurring in nature.

Simulated Annealing

In condensed-matter physics, the term 'annealing' refers to a physical process to obtain a pure lattice structure, where a solid is first heated up in a heat bath until it melts, and next cooled down slowly until it solidifies into a low-energy state. During the process, the free energy of the system is minimized. Simulated annealing, introduced in 1983 by S. Kirkpatrick, C.D. Gelatt and M.P. Vecchi [65], is a randomized neighborhood search algorithm based on the physical annealing process. Here, the solutions of a combinatorial optimization problem correspond to the states of the physical system, and the cost of a solution is equivalent to the energy of the state.

In its original formulation, simulated annealing works essentially as follows. Initially, a tentative solution in the state space is somehow generated. A new neighboring state is then produced from the previous one and, if the value of the cost function f improves, the new state is accepted, otherwise it is accepted with probability $\exp\{\Delta f/\tau\}$, where Δf is the difference of the cost function between the new and the current state, and τ is a parameter usually called the *temperature* in

analogy with physical annealing, which is varied carefully during the optimization process. The algorithm proceeds iteratively this way until a stopping condition is met. One of the critical aspects of the algorithm relates to the choice of the proper 'cooling schedule,' i. e., how to decrease the temperature as the process evolves. While a logarithmic slow cooling schedule (yielding an exponential time algorithm) provably guarantees the exact solution, faster cooling schedules, producing acceptably good results, are in widespread use. Introductory textbooks describing both theoretical and practical issues of the algorithm are [1,66].

E. Aarts and J. Korst [1], without presenting any experimental result, suggested the use of simulated annealing for solving the independent set problem, using a *penalty function* approach. Here, the solution space is the set of all possible subsets of vertices of the graph G, and the problem is formulated as one of maximizing the cost function $f(V') = |V'| - \lambda\ |E'|$, where $|E'|$ is the number of edges in $G(V')$, and λ is a weighting factor exceeding 1.

M. Jerrum [61] conducted a theoretical analysis of the performance of a clique-finding *Metropolis process*, i. e., simulated annealing at fixed temperature, on random graphs. He proved that the expected time for the algorithm to find a clique that is only slightly bigger than that produced by a naive greedy heuristic grows faster than any polynomial in the number of vertices. This suggests that 'true' simulated annealing would be ineffective for the maximum clique problem.

Jerrum's conclusion seems to be contradicted by practical experience. In [56], S. Homer and M. Peinado compare the performance of three heuristics, namely the greedy heuristic developed in [62], a randomized version of the Boppana–Halldórsson subgraph-exclusion algorithm [24], and simulated annealing, over very large graphs. The simulated annealing algorithm was essentially that proposed by Aarts and Korst, with a simple cooling schedule. This penalty function approach was found to work better than the method in which only cliques are considered, as proposed in [61]. The algorithms were tested on various random graphs as well as on DIMACS benchmark graphs. The authors ran the algorithms over an SGI workstation for graphs with up to 10,000 vertices, and on a Connection Machine for graphs with up to 70,000 vertices. The overall conclusion was that simulated annealing outperforms the other competing algorithms; it also ranked among the best heuristics for maximum clique presented at the 1993 DIMACS challenge [63].

Neural Networks

Artificial neural networks (often simply referred to as 'neural networks') are massively parallel, distributed systems inspired by the anatomy and physiology of the cerebral cortex, which exhibit a number of useful properties such as learning and adaptation, universal approximation, and pattern recognition (see [50,52] for an introduction).

In the mid 1980s, J.J. Hopfield and D.W. Tank [57] showed that certain feedback continuous neural models are capable of finding approximate solutions to difficult optimization problems such as the traveling salesman problem [57]. This application was motivated by the property that the temporal evolution of these models is governed by a quadratic Liapunov function (typically called 'energy function' because of its analogy with physical systems) which is iteratively minimized as the process evolves. Since then, a variety of combinatorial optimization problems have been tackled within this framework. The customary approach is to formulate the original problem as one of energy minimization, and then to use a proper relaxation network to find minimizers of this function. Almost invariably, the algorithms developed so far incorporate techniques borrowed from statistical mechanics, in particular mean field theory, which allow one to escape from poor local solutions. We mention the articles [69,82] and the textbook [88] for surveys of this field. In [1], an excellent introduction to a particular class of neural networks (the Boltzmann machine) for combinatorial optimization is provided.

Early attempts at encoding the maximum clique and related problems in terms of a neural network were already done in the late 1980s in [1,12,44,83], and [84] (see also [85]). However, little or no experimental results were presented, thereby making it difficult to evaluate the merits of these algorithms. In [68], F. Lin and K. Lee used the quadratic zero-one formulation from [78] as the basis for their neural network heuristic. On random graphs with up to 300 vertices, they found their algorithm to be faster than the implicit enumerative algorithm in [26], while obtaining slightly worse results in terms of clique size.

T. Grossman [45] proposed a discrete, deterministic version of the Hopfield model for maximum clique, originally designed for an all-optical implementation. The model has a threshold parameter which determines the character of the stable states of the network. The author suggests an annealing strategy on this parameter, and an adaptive procedure to choose the network's initial state and threshold. On DIMACS graphs the algorithm performs satisfactorily but it does not compare well with more powerful heuristics such as simulated annealing.

A. Jagota [58] developed several variations of the Hopfield model, both discrete and continuous, to approximate maximum clique. He evaluated the performance of his algorithms over randomly generated graphs as well as on harder graphs obtained by generating cliques of varying size at random and taking their union. Experiments on graphs coming from the Solomonoff–Levin, or 'universal' distribution are also presented in [59]. The best results were obtained using a stochastic steepest descent dynamics and a mean-field annealing algorithm, an efficient deterministic approximation of simulated annealing. These algorithms, however, were also the slowest, and this motivated Jagota et al. [60] to improve their running time. The mean-field annealing heuristic was implemented on a 32-processor Connection Machine, and a two-temperature annealing strategy was used. Additionally, a 'reinforcement learning' strategy was developed for the stochastic steepest descent heuristic, to automatically adjust its internal parameters as the process evolves. On various benchmark graphs, all their algorithms obtained significantly larger cliques than other simpler heuristics but ran slightly slower. Compared to more sophisticated heuristics, they obtained significantly smaller cliques on average but were considerably faster.

M. Pelillo [80] takes a completely different approach to the problem, by exploiting a continuous formulation of maximum clique and the dynamical properties of the so-called relaxation labeling networks. His algorithm is described in the next section.

Genetic Algorithms

Genetic algorithms are parallel search procedures inspired from the mechanisms of evolution in natural systems [45,55]. In contrast to more traditional optimization techniques, they work on a population of points, which in the genetic algorithm terminology, are called chromosomes or individuals. In the simplest and most popular implementation, chromosomes are simply long strings of bits. Each individual has an associated 'fitness' value which determines its probability of survival in the next 'generation': the higher the fitness, the higher the probability of survival. The genetic algorithm starts out with an initial population of members generally chosen at random and, in its simplest version, makes use of three basic operators: reproduction, crossover and mutation. Reproduction usually consists of choosing the chromosomes to be copied in the next generation according to a probability proportional to their fitness. After reproduction, the crossover operator is applied between pairs of selected individuals to produce new offsprings. The operator consists of swapping two or more subsegments of the the strings corresponding to the two chosen individuals. Finally, the mutation operator is applied, which randomly reverses the value of every bit within a chromosome with a fixed probability. The procedure just described is sometimes referred to as the 'simple' genetic algorithm [45].

One of the earliest attempts to solve the maximum clique problem using genetic algorithms was done in 1993 by B. Carter and K. Park [27]. After showing the weakness of the simple genetic algorithm in finding large cliques, even on small random graphs, they introduced several modifications in an attempt to improve performance. However, despite their efforts they did not get satisfactory results, and their general conclusion was that genetic algorithms need to be heavily customized in order to be competitive with traditional approaches, and that they are computationally very expensive. In a later study [79], genetic algorithms were proven to be less effective than simulated annealing. At almost the same time, T. Bäck and S. Khuri [8], working on the maximum independent set problem, arrived at the opposite conclusion. By using a straightforward, general-purpose genetic algorithm called GENEsYs and a suitable fitness function which included a graded penalty term to penalize infeasible solutions, they got interesting results over random and regular graphs with up to 200 vertices. These results indicate that the choice of the fitness function is crucial for genetic algorithms to provide satisfactory results.

A.S. Murthy et al. [74] also experimented with a genetic algorithm using a novel 'partial copy crossover', and a modified mutation operator. However, they presented results over very small (i. e., up to 50 vertices) graphs, thereby making it difficult to properly evaluate the algorithm.

T.N. Bui and P.H. Eppley [25] obtained encouraging results by using a hybrid strategy which incorporates a local optimization step at each generation of the genetic algorithm, and a vertex-ordering preprocessing phase. They tested the algorithm over some DIMACS graphs getting results comparable to that in[39]

Instead of using the standard binary representation for chromosomes, J.A. Foster and T. Soule [36] employed an integer-based encoding scheme. Moreover, they used a time weighting fitness function similar in spirit to those in [27]. The results obtained are interesting, but still not comparable to those obtained using more traditional search heuristics.

C. Fleurent and J.A. Ferland [35] developed a general-purpose system for solving graph coloring, maximum clique, and satisfiability problems. As far as the maximum clique problem is concerned, they conducted several experiments using a hybrid genetic search scheme which incorporates tabu search and other local search techniques as alternative mutation operators. The results presented are encouraging, but running time is quite high.

In [53], M. Hifi modifies the basic genetic algorithm in several aspects:

a) a particular crossover operator creates two new different children;
b) the mutation operator is replaced by a specific heuristic feasibility transition adapted to the weighted maximum stable set problem.

This approach is also easily parallelizable. Experimental results on randomly generated graphs and also some (unweighted) instances from the DIMACS testbed [63] are reported to validate this approach.

Finally, E. Marchiori [71] has developed a simple heuristic-based genetic algorithm which consists of a combination of the simple genetic algorithm and a naive greedy heuristic procedure. Unlike previous approaches, here there is a neat division of labor, the search for a large subgraph and the search for a clique being incorporated into the fitness function and the heuristic procedure, respectively. The algorithm outperforms previous genetic-based clique finding procedures over various DIMACS graphs, both in terms of quality of solutions and speed.

Tabu Search

Tabu search, introduced independently by F. Glover [41,42] and P. Hansen and B. Jaumard [48], is a modified local search algorithm, in which a prohibition-based strategy is employed to avoid cycles in the search trajectories and to explore new regions in the search space. At each step of the algorithm, the next solution visited is always chosen to be the best *legal neighbor* of the current state, even if its cost is worse than the current solution. The set of legal neighbors is restricted by one or more *tabu lists* which prevent the algorithm to go back to recently visited solutions. These lists are used to store historical information on the path followed by the search procedure. Sometimes the tabu restriction is relaxed, and tabu solutions are accepted if they satisfy some *aspiration level* condition. The standard example of a tabu list is one which contains the last k solutions examined, where k may be fixed or variable. Additional lists containing the last modifications performed, i. e., changes occurred when moving from one solution to the next, are also common. These types of lists are referred to as *short-term memories*; other forms of memories are also used to intensify the search in a promising region or to diversify the search to unexplored areas. Details on the algorithm and its variants can be found in [43] and [51].

In 1989, C. Friden et al. [37] proposed a heuristic for the maximum independent set problem based on tabu search. The size of the independent set to search for is fixed, and the algorithm tries to minimize the number of edges in the current subset of vertices. They used three tabu lists: one for storing the last visited solutions and the other two to contain the last introduced/deleted vertices. They showed that by using hashing for implementing the first list and choosing a small value for the dimensions of the other two lists, a best neighbor may be found in almost constant time.

In [38,86], three variants of tabu search for maximum clique are presented. Here the search space consists of complete subgraphs whose size has to be maximized. The first two versions are deterministic algorithms in which no sampling of the neighborhood is

performed. The main difference between the two algorithms is that the first one uses just one tabu list (of the last solutions visited), while the second one uses an additional list (with an associated aspiration mechanism) containing the last vertices deleted. Also, two diversification strategies were implemented. The third algorithm is probabilistic in nature, and uses the same two tabu lists and aspiration mechanism as the second one. It differs from it because it performs a random sampling of the neighborhood, and also because it allows for multiple deletion of vertices in the current solution. Here no diversification strategy was used. In [38,86] results on randomly generated graphs were presented and the algorithms were shown to be very effective. P. Soriano and M. Gendreau [87] tested their algorithms over the DIMACS benchmark graphs and the results confirmed the early conclusions.

R. Battiti and M. Protasi [13] extended the tabu search framework by introducing a reactive local search method. They modified a previously introduced reactive scheme by exploiting the particular neighborhood structure of the maximum clique problem. In general reactive schemes aim at avoiding the manual selection of control parameters by means of an internal feedback loop. Battiti and Protasi's algorithm adopts such a strategy to automatically determine the so-called prohibition parameter k, i. e., the size of the tabu list. Also an explicit memory-influenced restart procedure is activated periodically to introduce diversification. The search space consists of all possible cliques, as in the approach by Friden et al., and the function to be maximized is the clique size. The worst-case computational complexity of this algorithm is $O(\max\{n, m\})$, where n and m are the number of vertices and edges of the graph respectively. They noticed, however, that in practice, the number of operations tends to be proportional to the average degree of the vertices of the complement graph. They tested their algorithm over many DIMACS benchmark graphs obtaining better results then those presented at the DIMACS workshop in competitive time.

Continuous Based Heuristics

In 1965, T.S. Motzkin and E.G. Straus [73] established a remarkable connection between the maximum clique problem and a certain quadratic programming problem. Let $G = (V, E)$ be an undirected (unweighted) graph and let Δ denote the standard simplex in the n-dimensional Euclidean space $\mathbf{R}^n$:

$$\Delta = \left\{x \in \mathbf{R}^n : \; x_i \geq 0 \text{ for all } i \in V, \; e^\top x = 1\right\},$$

where the letter e is reserved for a vector of appropriate length, consisting of unit entries (hence $e^\top x = \sum_{i \in V} x_i$).

Now, consider the following quadratic function, sometimes called the *Lagrangian* of G:

$$g(x) = x^\top A_G x, \tag{1}$$

where $A_G = (a_{ij})$ is the adjacency matrix of G, i. e. the symmetric $n \times n$ matrix where $a_{ij} = 1$ if $(i, j) \in E$ and $a_{ij} = 0$ if $(i, j) \notin E$, and let x^* be a global maximizer of g on Δ. In [73] it is proved that the clique number of G is related to $g(x^*)$ by the following formula:

$$\omega(G) = \frac{1}{1 - g(x^*)}.$$

Additionally, it is shown that a subset of vertices S is a maximum clique of G if and only if its *characteristic vector* x^S, which is the vector of Δ defined as $x_i^S = 1/|S|$ if $i \in S$ and $x_i^S = 0$ otherwise, is a global maximizer of g on Δ. In [40,81], the Motzkin–Straus theorem has been extended by providing a characterization of maximal cliques in terms of local maximizers of g on Δ.

One drawback associated with the original Motzkin–Straus formulation relates to the existence of spurious solutions, i. e., maximizers of g which are not in the form of characteristic vectors [77,81]. In principle, spurious solutions represent a problem since, while providing information about the cardinality of the maximum clique, they do not allow us to easily extract its vertices.

During the 1990s, there has been much interest around the Motzkin–Straus and related continuous formulations of the maximum clique problem. They suggest in fact a fundamentally new way of solving the maximum clique problem, by allowing us to shift from the discrete to the continuous domain in an elegant manner. As pointed out in [76], continuous formulations of discrete optimization problems turn out to be particularly attractive. They not only allow us to exploit the full arsenal of continuous optimization techniques, thereby leading to the development of new algorithms, but may also reveal unexpected theoretical properties.

In [77], P.M. Pardalos and A.T. Phillips developed a global optimization approach based on the Motzkin–Straus formulation and implemented an iterative clique retrieval process to find the vertices of the maximum clique. However, due to its high computational cost they were not able to run the algorithm over graphs with more than 75 vertices.

Pelillo [80] used *relaxation labeling algorithms* to approximately determining the size of the maximum clique using the original Motzkin–Straus formulation. These are parallel, distributed algorithms developed and studied in computer vision and pattern recognition, which are also surprisingly related to *replicator equations*, a class of dynamical systems widely studied in evolutionary game theory and related fields [54], Heuristics for maximum clique and independent set. The model operates in the simplex Δ and possesses a quadratic Liapunov function which drives its dynamical behavior. It is these properties that naturally suggest using them as a local optimization algorithm for the Motzkin–Straus program. The algorithm is especially suited for parallel implementation, and is attractive for its operational simplicity, since no parameters need to be determined. Extensive simulations over random graphs with up to 2000 vertices have demonstrated the effectiveness of the approach and showed that the algorithm outperforms previous neural network heuristics.

In order to avoid time-consuming iterative procedures to extract the vertices of the clique, L.E. Gibbons, D.W. Hearn and Pardalos [39] have proposed a heuristic which is based on a parameterized formulation of the Motzkin–Straus program. They consider the problem of minimizing the function:

$$h(x) = \frac{1}{2}x^\top A_{\overline{G}}x + \left(\sum_{i=1}^{n} x_i - 1\right)^2$$

on the domain:

$$S(k) = \left\{ x \in \mathbf{R}^n : \ \sum_{i=1}^{n} x_i^2 \le \frac{1}{k},\ x_i \ge 0,\ \forall i \right\},$$

where k is a fixed parameter. Let x^* be a global minimizer of h on $S(k)$, and let $V(k) = h(x^*)$. In [39] it is proved that $V(k) = 0$ if and only if there exists an independent set S of $\overline{G}$ with size $|S| \ge k$. Moreover, the vertices of $\overline{G}$ associated with the indices of the positive components of x^* form an independent set of size greater than or equal k.

These properties motivated the following procedure to find a maximum independent set of $\overline{G}$ or, equivalently, a maximum clique of G. Minimize the function h over $S(k)$, for different values of k between predetermined upper and lower bounds. If $V(k) = 0$ and $V(k+1) \neq 0$ for some k, then the maximum clique of G has size k, and its vertices are determined by the positive components of the solution. Since minimizing h on $S(k)$ is a difficult problem, they developed a heuristic based on the observation that by removing the nonnegativity constraints, the problem is that of minimizing a quadratic form over a sphere, a problem which is solvable in polynomial time. However, in so doing a heuristic procedure is needed to round the approximate solutions of this new problem to approximate solutions of the original one. Moreover, since the problem is solved approximately, we have to find the value of the spherical constraint $1k$ which yields the largest independent set. A careful choice of k is therefore needed. The resulting algorithm was tested over various DIMACS benchmark graphs [63] and the results obtained confirmed the effectiveness of the approach.

The spurious solution problem has been solved in [16]. Consider the following regularized version of function g:

$$\widehat{g}(x) = x^\top A_G x + \frac{1}{2}x^\top x \tag{2}$$

which is obtained from (1) by substituting the adjacency matrix A_G of G with

$$\widehat{A}_G = A_G + \frac{1}{2}I,$$

where I is the identity matrix. Unlike the Motzkin–Straus formulation, it can be proved that all maximizers of $\widehat{g}$ on Δ are strict, and are characteristic vectors of maximal/maximum cliques in the graph. In an exact sense, therefore, a one-to-one correspondence exists between maximal cliques and local maximizers of $\widehat{g}$ in Δ on the one hand and maximum cliques and global maximizers on the other hand. In [16,20], replicator equations are used in conjunction to this spurious-free formulation to find maximal cliques of G. Note that here the vertices comprising the clique are directly given by the positive components of the converged vectors, and no iterative procedure is needed to determine

them, as in [77]. The results obtained over a set of random as well as DIMACS benchmark graphs were encouraging, especially considering that replicator equations do not incorporate any mechanism to escape from local optimal solutions. This suggests that the basins of attraction of the global solution with respect to the quadratic functions g and $\widehat{g}$ are quite large; for a thorough empirical analysis see also [23]. One may wonder whether a subtle choice of initial conditions and/or a variant of the dynamics may significantly improve the results, but experiments in [22] indicate this is not the case.

In [19] the properties of the following function are studied:

$$\widehat{g}_\alpha(x) = x^\top A_G x + \alpha x^\top x.$$

It is shown that when α is positive all the properties enjoyed by the standard regularization approach [16] hold true. Specifically, in this case a one-to-one correspondence between local/global maximizers in the continuous space and local/global solutions in the discrete space exists. For negative α's an interesting picture emerges: as the absolute value of α grows larger, local maximizers corresponding to maximal cliques disappear. In [19], bounds on the parameter α are derived which affect the stability of these solutions. These results have suggested an *annealed replication heuristic*, which consists of starting from a large negative α and then properly reducing it during the optimization process. For each value of α standard replicator equations are run in order to obtain local solutions of the corresponding objective function. The rationale behind this idea is that for values of α with a proper large absolute value only local solutions corresponding to large maximal cliques will survive, together with various spurious maximizers. As the value of α is reduced, spurious solutions disappear and smaller maximal cliques become stable. An annealing schedule is proposed which is based on the assumption that the graphs being considered are random. In this respect, the proposed procedure differs from usual simulated annealing approaches, which mostly use a 'black-box' cooling schedule. Experiments conducted over several DIMACS benchmark graphs confirm the effectiveness of the proposed approach and the robustness of the annealing strategy. The overall conclusion is that the annealing procedure does help to avoid inefficient local solutions, by initially driving the dynamics towards promising regions in state space, and then refining the search as the annealing parameter is increased.

The Motzkin–Straus theorem has been generalized to the weighted case in [40]. Note that the Motzkin–Straus program can be reformulated as a minimization problem by considering the function

$$f(x) = x^\top (I + A_{\overline{G}})x,$$

where $A_{\overline{G}}$ is the adjacency matrix of the complement graph $\overline{G}$. It is straightforward to see that if x^* is a global minimizer of f in Δ, then we have: $\omega(G) = 1/\mathrm{f}(x^*)$. This is simply a different formulation of the Motzkin–Straus theorem. Given a weighted graph $G = (V, E)$ with weight vector w, let $\mathcal{M}(G, w)$ be the class of symmetric $n \times n$ matrices $B = (b_{ij})_{i,j \in V}$ defined as $2b_{ij} \geq b_{ii} + b_{jj}$ if $(i, j) \notin E$ and $b_{ij} = 0$ otherwise, and $b_{ii} = 1/w_i$ for all $i \in V$.

Given the following quadratic program, which is in general indefinite,

$$\begin{cases} \min & f(x) = x^\top Bx \\ \text{s.t.} & x \in \Delta, \end{cases} \tag{3}$$

in [40] it is shown that for any $B \in \mathcal{M}(G, w)$ we have:

$$\omega(G, w) = \frac{1}{f(x^*)},$$

where x^* is a global minimizer of program (3). Furthermore, denote by x^S the *weighted characteristic vector* of S, which is a vector with coordinates $x_i^S = w_i/W(S)$ if $i \in S$ and $x_i^S = 0$ otherwise. It can be seen that a subset S of vertices of a weighted graph G is a maximum weight clique if and only if its characteristic vector x^S is a global minimizer of (3). Notice that the matrix $I + A_{\overline{G}}$ belongs to $\mathcal{M}(G, e)$. In other words, the Motzkin–Straus theorem turns out to be a special case of the preceding result.

As in the unweighted case, the existence of spurious solutions entails the lack of one-to-one correspondence between the solutions of the continuous problem and those of the original, discrete one. In [21] these spurious solutions are characterized and a regularized version which avoids this kind of problems is introduced, exactly as in the unweighted case (see also [17]). Replicator equations are then used to find maximal weight

cliques in weighted graphs, using this formulation. Experiments with this approach on both random graphs and DIMACS graphs are reported. The results obtained are compared with those produced by a very efficient maximum weight clique algorithm of the branch and bound variety. The algorithm performed remarkably well especially on large and dense graphs, and it was typically an order of magnitude more efficient than its competitor.

Finally, we mention the work by Massaro and Pelillo [72], who transformed the Motzkin–Straus program into a linear complementarity problem [31], and then solved it using a variation of Lemke's well-known algorithm [67]. The preliminary results obtained over many weighted and unweighted DIMACS graphs show that this approach substantially outperforms all other continuous based heuristics.

Miscellaneous

Another type of heuristics that finds a maximal clique of G is called the *subgraph approach* (see [11]). It is based on the fact that a maximum clique C of a subgraph $G' \subseteq G$ is also a clique of G. The subgraph approach first finds a subgraph $G' \subseteq G$ such that the maximum clique of G' can be found in polynomial time. Then it finds a maximum clique of G' and use it as the approximation solution. The advantage of this approach is that in finding the maximum clique $C \subseteq G'$, one has (implicitly) searched many other cliques of G' ($C_{G'} \subseteq C_G$). Because of the special structure of G', this implicit search can be done efficiently. In [11], G' is a maximal induced triangulated subgraph of G. Since many classes of graphs have polynomial algorithms for the maximum clique problem, the same idea also applies there. For example, the class of edge-maximal triangulated subgraphs was chosen in [9,90], and [91]. Some of the greedy heuristics, randomized heuristics and subgraph approach heuristics are compared in [90] and [91] on randomly generated weighted and unweighted graphs.

Various new heuristics were presented at the 1993 DIMACS challenge devoted to clique, coloring and satisfiability [63]. In particular, in [10] an algorithm is proposed which is based on the observation that finding the maximum clique in the union of two cliques can be done using bipartite matching techniques. In [46] restricted backtracking is used to provide a trade-off between the size of the clique and the completeness of the search. In [70] an edge projection technique is proposed to obtain a new upper bound heuristic for the maximum independent set problem. This procedure was used, in conjunction with the Balas–Yu branching rule [11], to develop an exact branch and bound algorithm which works well especially on sparse graphs.

See [3] for a new population-based optimization heuristic inspired by the natural behavior of human or animal scouts in exploring unknown regions, and applied it to maximum clique. The results obtained over a few DIMACS graphs are comparable with those obtained using continuous-based heuristics but are inferior to those obtained by reactive local search.

Recently, DNA computing [4] has also emerged as a potential technique for the maximum clique problem [75,92]. The major advantage of DNA computing is its high parallelism, but at present the size of graphs this algorithm can handle is limited to a few tens.

Additional heuristics for the maximum clique/independent set and related problems on arbitrary or special class of graphs can be found in [28,29,30,34].

Conclusions

During the 1990s, research on the maximum clique and related problems has yielded many interesting heuristics. This article has provided an expository survey on these algorithms and an up-to-date bibliography (as of 2000). However, the activity in this field is so extensive that a survey of this nature is outdated before it is written.

See also

- ► Graph Coloring
- ► Greedy Randomized Adaptive Search Procedures
- ► Replicator Dynamics in Combinatorial Optimization

References

1. Aarts E, Korst J (1989) Simulated annealing and Boltzmann machines. Wiley, New York
2. Aarts E, Lenstra JK (eds) (1997) Local search in combinatorial optimization. Wiley, New York
3. Abbattista F, Bellifemmine F, Dalbis D (1998) The Scout algorithm applied to the maximum clique problem. Ad-

vances in Soft Computing—Engineering and Design. Springer, Berlin
4. Adleman LM (1994) Molecular computation of solutions to combinatorial optimization. Science 266:1021–1024
5. Alon N, Babai L, Itai A (1986) A fast and simple randomized parallel algorithm for the maximal independent set problem. J Algorithms 7:567–583
6. Arora S, Lund C, Motwani R, Sudan M, Szegedy M (1992) Proof verification and the hardness of approximation problems. Proc. 33rd Ann. Symp. Found. Comput. Sci., Pittsburgh, pp 14–23
7. Arora S, Safra S (1992) Probabilistic checking of proofs: A new characterization of NP. Proc. 33rd Ann. Symp. Found. Comput. Sci., Pittsburgh, pp 2–13
8. Bäck T, Khuri S (1994) An evolutionary heuristic for the maximum independent set problem. Proc. 1st IEEE Conf. Evolutionary Comput., 531–535
9. Balas E (1986) A fast algorithm for finding an edge-maximal subgraph with a TR-formative coloring. Discrete Appl Math 15:123–134
10. Balas E, Niehaus W (1996) Finding large cliques in arbitrary graphs by bipartite matching. In: Johnson DS, Trick MA (eds) Cliques, Coloring, and Satisfiability: 2nd DIMACS Implementation Challenge, DIMACS 26. AMS, Providence, RI, pp 29–51
11. Balas E, Yu CS (1986) Finding a maximum clique in an arbitrary graph. SIAM J Comput 14:1054–1068
12. Ballard DH, Gardner PC, Srinivas MA (1987) Graph problems and connectionist architectures. Techn Report Dept Comput Sci Univ Rochester 167
13. Battiti R, Protasi M (1995) Reactive local search for the maximum clique problem. TR-95-052 Techn Report Internat Comput Sci Inst Berkeley, to appear in Algorithmica
14. Bellare M, Goldwasser S, Lund C, Russell A (1993) Efficient probabilistically checkable proofs and application to approximation. Proc. 25th Ann. ACM Symp. Theory Comput., pp 294–304
15. Bellare M, Goldwasser S, Sudan M (1995) Free bits, PCPs and non-approximability – Towards tight results. Proc. 36th Ann. Symp. Found. Comput. Sci., pp 422–431
16. Bomze IM (1997) Evolution towards the maximum clique. J Global Optim 10:143–164
17. Bomze IM (1998) On standard quadratic optimization problems. J Global Optim 13:369–387
18. Bomze IM, Budinich M, Pardalos PM, Pelillo M (1999) The maximum clique problem. In: Du D-Z, Pardalos PM (eds) Handbook Combinatorial Optim., Suppl. A. Kluwer, Dordrecht, pp 1–74
19. Bomze IM, Budinich M, Pelillo M, Rossi C (2000) An new "annealed" heuristic for the maximum clique problem. In: Pardalos PM (eds) Approximation and Complexity in Numerical Optimization: Continuous and Discrete Problems. Kluwer, Dordrecht, pp 78–95
20. Bomze IM, Pelillo M, Giacomini R (1997) Evolutionary approach to the maximum clique problem: Empirical evidence on a larger scale. In: Bomze IM, Csendes T, Horst R, Pardalos PM (eds) Developments in Global Optimization. Kluwer, Dordrecht, pp 95–108
21. Bomze IM, Pelillo M, Stix V (2000) Approximating the maximum weight clique using replicator dynamics. IEEE Trans Neural Networks 11(6)
22. Bomze IM, Rendl F (1998) Replicator dynamics for evolution towards the maximum clique: Variations and experiments. In: De Leone R, Murli A, Pardalos PM, Toraldo G (eds) High Performance Algorithms and Software in Nonlinear Optimization. Kluwer, Dordrecht, pp 53–67
23. Bomze IM, Stix V (1999) Genetic engineering via negative fitness: Evolutionary dynamics for global optimization. Ann Oper Res 90
24. Boppana R, Halldóorsson MM (1992) Approximating maximum independent sets by excluding subgraphs. BIT 32:180–196
25. Bui TN, Eppley PH (1995) A hybrid genetic algorithm for the maximum clique problem. Proc. 6th Internat. Conf. Genetic Algorithms, pp 478–484
26. Carraghan R, Pardalos PM (1990) An exact algorithm for the maximum clique problem. Oper Res Lett 9:375–382
27. Carter B, Park K (1993) How good are genetic algorithms at finding large cliques: An experimental study. Techn Report Comput Sci Dept Boston Univ no. BU-CS-93-015
28. Chiba N, Nishizeki T, Saito N (1983) An algorithm for finding a large independent set in planar graphs. Networks 13:247–252
29. Chrobak M, Naor J (1991) An efficient parallel algorithm for computing a large independent set in a planar graph. Algorithmica 6:801–815
30. Chvátal V (1979) A greedy heuristic for the set-covering problem. Math Oper Res 4:233–23
31. Cottle RW, Pang J, Stone RE (1992) The linear complementarity problem. AP
32. Feige U, Goldwasser S, Lováasz L, Safra S, Szegedy M (1991) Approximating clique is almost NP-complete. Proc. 32nd Ann. Symp. Found. Comput. Sci., San Juan, Puerto Rico, pp 2–12
33. Feo TA, Resende MGC, Smith SH (1994) A greedy randomized adaptive search procedure for maximum independent set. Oper Res 42:860–878
34. Fisher ML, Wolsey LA (1982) On the greedy heuristic for continuous covering and packing problems. SIAM J Alg Discrete Meth 3:584–591
35. Fleurent C, Ferland JA (1996) Object-oriented implementation of heuristic search methods for graph coloring, maximum clique, and satisfiability. In: Johnson DS, Trick MA (eds) Cliques, Coloring, and Satisfiability: 2nd DIMACS Implementation Challenge, DIMACS 26. AMS, Providence, RI, pp 619–652
36. Foster JA, Soule T (1995) Using genetic algorithms to find maximum cliques. Techn Report Dept Comput Sci Univ Idaho no. LAL 95-12
37. Friden C, Hertz A, de Werra M (1989) STABULUS: A tech-

nique for finding stable sets in large graphs with tabu search. Computing 42:35–44
38. Gendreau A, Salvail L, Soriano P (1993) Solving the maximum clique problem using a tabu search approach. Ann Oper Res 41:385–403
39. Gibbons LE, Hearn DW, Pardalos PM (1996) A continuous based heuristic for the maximum clique problem. In: Johnson DS, Trick MA (eds) Cliques, Coloring, and Satisfiability: 2nd DIMACS Implementation Challenge, DIMACS 26. AMS, Providence, RI, pp 103–124
40. Gibbons LE, Hearn DW, Pardalos PM, Ramana MV (1997) Continuous characterizations of the maximum clique problem. Math Oper Res 22:754–768
41. Glover F (1989) Tabu search–Part I. ORSA J Comput 1:190–260
42. Glover F (1990) Tabusearch–Part II. ORSA J Comput 2:4–32
43. Glover F, Laguna M (1993) Tabu search. In: Reeves C (ed) Modern Heuristic Techniques for Combinatorial Problems. Blackwell, Oxford pp 70–141
44. Godbeer GH, Lipscomb J, Luby M (1988) On the computational complexity of finding stable state vectors in connectionist models (Hopfield nets). Techn Report Dept Comput Sci Univ Toronto 208,
45. Goldberg DE (1989) Genetic algorithms in search, optimization, and machine learning. Addison-Wesley, Reading, MA
46. Goldberg MK, Rivenburgh RD (1996) Constructing cliques using restricted backtracking. In: Johnson DS, Trick MA (eds) Cliques, Coloring, and Satisfiability: 2nd DIMACS Implementation Challenge, DIMACS 26. AMS, Providence, RI, pp 89–101
47. Grossman T (1996) Applying the INN model to the max clique problem. In: Johnson DS, Trick MA (eds) Cliques, Coloring, and Satisfiability: 2nd DIMACS Implementation Challenge, DIMACS 26. AMS, Providence, RI, pp 125–146
48. Hansen P, Jaumard B (1990) Algorithms for the maximum satisfiability problem. Computing 44:279–303
49. Håstad J (1996) Clique is hard to approximate within $n^{1-\epsilon'}$:. Proc. 37th Ann. Symp. Found. Comput. Sci., pp 627–636
50. Haykin S (1994) Neural networks: A comprehensive foundation. MacMillan, New York
51. Hertz A, Taillard E, de Werra D (1997) Tabu search. In: Aarts E, Lenstra JK (eds) Local Search in Combinatorial Optimization. Wiley, New York pp 121–136
52. Hertz J, Krogh A, Palmer RG (1991) Introduction to the theory of neural computation. Addison-Wesley, Reading, MA
53. Hifi M (1997) A genetic algorithm-based heuristic for solving the weighted maximum independent set and some equivalent problems. J Oper Res Soc 48:612–622
54. Hofbauer J, Sigmund K (1998) Evolutionary games and population dYnamics. Cambridge Univ. Press, Cambridge
55. Holland JH (1975) Adaptation in natural and artificial systems. Univ. Michigan Press, Ann Arbor, MI
56. Homer S, Peinado M (1996) Experiments with polynomial-time CLIQUE approximation algorithms on very large graphs. In: Johnson DS, Trick MA (eds) Cliques, Coloring, and Satisfiability: 2nd DIMACS Implementation Challenge, DIMACS 26. AMS, Providence, RI, pp 147–167
57. Hopfield JJ, Tank DW (1985) Neural computation of decisions in optimization problems. Biol Cybern 52:141–152
58. Jagota A (1995) Approximating maximum clique with a Hopfield neural network. IEEE Trans Neural Networks 6:724–735
59. Jagota A, Regan KW (1997) Performance of neural net heuristics for maximum clique on diverse highly compressible graphs. J Global Optim 10:439–465
60. Jagota A, Sanchis L, Ganesan R (1996) Approximately solving maximum clique using neural networks and related heuristics. In: Johnson DS, Trick MA (eds) Cliques, Coloring, and Satisfiability: 2nd DIMACS Implementation Challenge, DIMACS 26. AMS, Providence, RI, pp 169–204
61. Jerrum M (1992) Large cliques elude the Metropolis process. Random Struct Algorithms 3:347–359
62. Johnson DS (1974) Approximation algorithms for combinatorial problems. J Comput Syst Sci 9:256–278
63. Johnson DS, Trick MA (eds) (1996) Cliques, coloring, and satisfiability: 2nd DIMACS implementation challenge, DIMACS 26. AMS, Providence, RI
64. Karp RM (1972) Reducibility among combinatorial problems. In: Miller RE, Thatcher JW (eds) Complexity of Computer Computations. Plenum, New York, pp 85–103
65. Kirkpatrick S, Gelatt CD, Vecchi MP (1983) Optimization by simulated annealing. Science 220:671–680
66. van Laarhoven PJM, Aarts EHL (1987) Simulated annealing: Theory and applications. Reidel, London
67. Lemke CE (1965) Bimatrix equilibrium points and mathematical programming. Managem Sci 11:681–689
68. Lin F, Lee K (1992) A parallel computation network for the maximum clique problem. Proc. 1st Internat. Conf. Fuzzy Theory Tech.,
69. Looi C-K (1992) Neural network methods in combinatorial optimization. Comput Oper Res 19:191–208
70. Mannino C, Sassano A (1996) Edge projection and the maximum cardinality stable set problem. In: Johnson DS, Trick MA (eds) Cliques, Coloring, and Satisfiability: 2nd DIMACS Implementation Challenge, DIMACS 26. AMS, Providence, RI, pp 205–219
71. Marchiori E (1998) A simple heuristic based genetic algorithm for the maximum clique problem. Proc. ACM Symp. Appl. Comput., pp 366–373
72. Massaro A, Pelillo M (2000) A pivoting-based heuristic for the maximum clique problem. Presented at the Internat. Conf. Advances in Convex Analysis and Global Optimization, Samos, June 2000
73. Motzkin TS, Straus EG (1965) Maxima for graphs and a new proof of a theorem of Turán. Canad J Math 17:533–540
74. Murthy AS, Parthasarathy G, Sastry VUK (1994) Clique finding–A genetic approach. Proc. 1st IEEE Conf. Evolutionary Comput., pp 18–21

75. Ouyang Q, Kaplan PD, Liu S, Libchaber A (1997) DNA solutions of the maximal clique problem. Science 278:1446–449
76. Pardalos PM (1996) Continuous approaches to discrete optimization problems. In: Di Pillo G, Giannessi F (eds) Nonlinear Optimization and Applications. Plenum, New York pp 313–328
77. Pardalos PM, Phillips AT (1990) A global optimization approach for solving the maximum clique problem. Internat J Comput Math 33:209–216
78. Pardalos PM, Rodgers GP (1990) Computational aspects of a branch and bound algorithm for quadratic zero-one programming. Computing 45:131–144
79. Park K, Carter B (1994) On the effectiveness of genetic search in combinatorial optimization. no.BU-CS-9-010 Techn Report Computer Sci Dept Boston Univ
80. Pelillo M (1995) Relaxation labeling networks for the maximum clique problem. J Artif Neural Networks 2:313–328
81. Pelillo M, Jagota A (1995) Feasible and infeasible maxima in a quadratic program for maximum clique. J Artif Neural Networks 2:411–420
82. Peterson C, Söderberg B (1997) Artificial neural networks. In: Aarts E, Lenstra JK (eds) Local Search in Combinatorial Optimization. Wiley, New York pp 173–213
83. Ramanujam J, Sadayappan P (1988) Optimization by neural networks. Proc. IEEE Internat. Conf. Neural Networks, pp 325–332
84. Shrivastava Y, Dasgupta S, Reddy SM (1990) Neural network solutions to a graph theoretic problem. Proc. IEEE Internat. Symp. Circuits Syst., pp 2528–2531
85. Shrivastava Y, Dasgupta S, Reddy SM (1992) Guaranteed convergence in a class of Hopfield networks. IEEE Trans Neural Networks 3:951–961
86. Soriano P, Gendreau M (1996) Diversification strategies in tabu search algorithms for the maximum clique problem. Ann Oper Res 63:189–207
87. Soriano P, Gendreau M (1996) Tabu search algorithms for the maximum clique problem. In: Johnson DS, Trick MA (eds) Cliques, Coloring, and Satisfiability: 2nd DIMACS Implementation Challenge, DIMACS 26. AMS, Providence, RI, pp 221–242
88. Takefuji Y (1992) Neural network parallel computing. Kluwer, Dordrecht
89. Tomita E, Mitsuma S, Takahashi H (1988) Two algorithms for finding a near-maximum clique. Techn Report UEC-TR-C1
90. Xue J (1991) Fast algorithms for vertex packing and related problems. PhD Thesis GSIA Carnegie-Mellon Univ.
91. Xue J (1994) Edge-maximal triangulated subgraphs and heuristics for the maximum clique problem. Networks 24:109–120
92. Zhang B-T, Shin S-Y (1998) Code optimization for DNA computing of maximal clique's. Advances in Soft Computing–Engineering and Design. Springer, Berlin

High-order Maximum Principle for Abnormal Extremals

Urszula Ledzewicz[1], Heinz Schättler[2]
[1] Department Math. and Statist., Southern Illinois University at Edwardsville, Edwardsville, USA
[2] Department Systems Sci. and Math., Washington University, St. Louis, USA

MSC2000: 49K15, 49K27, 41A10, 47N10

Article Outline

Keywords

Local maximum principle; High-order tangent sets; High-order necessary conditions for optimality; Abnormal processes

We formulate a generalized local maximum principle which gives necessary conditions for optimality of abnormal trajectories in optimal control problems. The results are based on a hierarchy of primal constructions of high-order approximating cones (consisting of tangent directions for equality constraints, feasible directions for inequality constraints, and directions of decrease for the objective) and dual characterizations of empty intersection properties of these cones (see ► High-order necessary conditions for optimality for abnormal points). Characteristic for the theorem is that the multiplier associated with the objective is nonzero.

We consider an optimal control problem in Bolza form with fixed terminal time:

(OC) Minimize the functional

$$I(x,u) = \int_0^T L(x(t),u(t),t)\,dt + \ell(x(T)) \qquad (1)$$

subject to the constraints

$$\dot{x}(t) = f(x(t), u(t), t),$$
$$x(0) = 0, \qquad q(x(T)) = 0,$$
$$u(\cdot) \in \mathcal{U} = \{u \in L^r_\infty(0, T):\ u(t) \in U\}.$$

The terminal time T is fixed and we make the following *regularity assumptions* on the data: L: $\mathbf{R}^n \times \mathbf{R}^m \times [0, T] \to \mathbf{R}$ and f: $\mathbf{R}^n \times \mathbf{R}^m \times \to [0, T]$ $\mathbf{R}^n$ are C_∞ in (x, u) for every $t \in [0, T]$; both functions and their derivatives are measurable in t for every (x, u) and the functions and all partial derivatives are bounded on compact subsets of $\mathbf{R}^n \times \mathbf{R}^m \times [0, T]$; ℓ: $\mathbf{R}^n \to \mathbf{R}$ and q: $\mathbf{R}^n \to \mathbf{R}^k$ are C^∞ and the rows of the Jacobian matrix q_x (i. e. the gradients of the equations defining the terminal constraint) are linearly independent; $U \subset \mathbf{R}^m$ is a closed and convex set with nonempty interior. Let

$$H(\lambda_0, \lambda, x, u, t) = \lambda_0 L(x, u, t) + \lambda f(x, u, t) \qquad (2)$$

be the Hamiltonian for the control problem. If the input-trajectory pair (x_*, u_*) is optimal for problem (OC), then the *local maximum principle* [7] states that there exist a constant $\lambda_0 \geq 0$, an absolutely continuous function λ:$[0, T] \to (\mathbf{R}^n)^*$ (which we write as a row vector), which is a solution to the adjoint equation

$$\dot{\lambda} = -H_x(\lambda_0, \lambda(t), x_*(t), u_*(t), t),$$

with terminal condition

$$\lambda(T) = \lambda_0 \ell_x(x_*(T)) + \nu q_x(x_*(T)), \qquad (3)$$

(for some row vector $\nu \in (\mathbf{R}^k)^*$) such that $(\lambda_0, \lambda(t)) \neq 0$ for all $t \in [0, T]$ and the following local minimum condition holds for all $u \in U$:

$$\langle H_u(\lambda_0, \lambda(t), x_*(t), u_*(t), t), u - u_*(t)\rangle \geq 0. \qquad (4)$$

Input-trajectory pairs (x_*, u_*) for which multipliers λ_0 and λ exist such that these conditions are satisfied are called (*weak*) *extremals*. If $\lambda_0 > 0$, then it is possible to normalize $\lambda_0 = 1$ and the extremal is called *normal* while extremals with $\lambda_0 = 0$ are called *abnormal*. Although the terminology abnormal, which has its origins in the calculus of variations [4], seems to suggest that these type of extremals are an aberration, for optimal control problems this is not the case. The phenomenon is quite general and abnormal extremals cannot be excluded from optimality a priori. For instance, there exist optimal abnormal trajectories for the standard problem of stabilizing the harmonic oscillator time-optimally in minimum time, a simple time-invariant linear system.

In the abnormal case conventional necessary conditions for optimality provide conditions which only describe the structure of the constraints. For example, if there are no control constraints, then these conditions only involve the equality constraint defined by the dynamics and terminal conditions as zero set of an operator F: $Z \to Y$ between Banach spaces. If $F'(z_*)$ is not onto, but $\mathrm{Im}F'(z_*)$ is closed (and this is always the case for the optimal control problem) then the standard Lagrange multiplier type necessary conditions for optimality (which imply the local maximum principle [7]) can be satisfied trivially by choosing a multiplier which annihilates the image of $F'(z_*)$ and setting all other multipliers to zero.) The corresponding necessary conditions are independent of the objective and describe only the structure of the constraint yielding little information about the optimality of the abnormal trajectory.

Much of the difficulty in analyzing abnormal points in extremum problems can be traced back to the fact that the equality constraint is typically no longer a manifold near abnormal points, but intersections of manifolds are common. Hence, in order to develop necessary and/or sufficient conditions for optimality of abnormal extremals, it is imperative to analyze different branches of the zero-set of F. Finding these branches is at the heart of the matter. Generalizing a result of E.R. Avakov [2,3] in [10] a high-order generalization of the *classical Lyusternik theorem* is given which for general $p \in \mathbf{N}$ describes the structure of p-order tangent directions to an operator equality constraint in a Banach space for nonregular operators under a more general surjectivity assumption involving the first p derivatives of the operator. Based on these results p-order tangent cones to the equality constraint can explicitly be calculated along critical directions which comprise the low order terms. Combining these cones with standard constructions of high-order cones of decrease for the functional and high-order feasible cones to inequality constraints, all taken along critical directions, generalized necessary conditions for optimality for extremum problems in Banach spaces can be derived which allow to incorporate the objective with a nonzero mul-

tiplier. Characteristic of these results is that they are parametrized by critical directions as it is 'natural' near abnormal points.

In [12] (see ► High-order necessary conditions for optimality for abnormal points) an abstract formulation of these results is presented for minimization problems in Banach spaces. The main result gives a dual characterization for the empty intersection property of the various approximating cones along critical directions, but primal arguments using the cones themselves are often equally effective. In this article we formulate these abstract results for the optimal control problem, but we only consider the so-called weak or local version of the maximum principle. This result is weaker than the *Pontryagin maximum principle* [15] in the sense that the Pontryagin maximum principle asserts that the Hamiltonian of the control problem is indeed minimized over the control set at every time along the reference trajectory by the reference control. The local version only gives the necessary conditions for optimality for this property. However, it is well-known how to use an argument of A. Ya. Dubovitskii to derive the Pontryagin maximum principle from the local version [7, Lecture 13] and a preliminary strong version of the results of this article is given in [9].

Other theories of necessary conditions which are tailored to abnormal processes include a method known as 'weakening equality constraints' introduced in [14] and developed further in [5]. References [2,3] are along the lines of the results described here and give necessary conditions for optimality of abnormal extremals based on quadratic approximations. Similarly, both weak and strong versions of a second order generalized maximum principle are given by the authors in [8]. While mostly optimization related techniques are used in these papers, on a different level [1] uses differential geometric techniques to develop a theory of the second variation for abnormal extremals. They give both necessary and sufficient conditions for so-called corank-1 abnormal extremals (extremals for which there exists a unique multiplier) in terms of the Jacobi equation and related Morse indices and nullity theorems. Second order necessary conditions for optimality in the type of accessory problem results without normality assumptions have first been given in [6]. Also, the results in [16] are derived without making normality assumptions.

Regularity in the Equality Constraint

We model the optimal control problem (OC) in the framework of optimization theory as a minimization problem in a Banach space under equality and inequality constraints. Let $W^n_{11}(0,\,T)$ denote the Banach space of all absolutely continuous functions $x\colon [0,\,T]\to \mathbf{R}^n$ with norm $|x| = \|x(0)\| + \int_0^T \|\dot{x}(s)\|\ ds$ and let $\overline{W}^n_{11}(0,\,T) = W^n_{11}(0,\,T)\cap\{x\in W^n_{11}(0,\,T)\colon\ x(0)=0\}$. Then the problem is to minimize the functional I over a class $\mathcal{A}$ of input-trajectory pairs $(x,u)\in \overline{W}^n_{11}(0,\,T)\times L^m_\infty(0,\,T)$ which is defined by equality constraints and the convex inequality constraint $u\in\mathcal{U}$. The equality constraints can be modeled as $\mathcal{F} = \{(x,u)\in \overline{W}^n_{11}(0,\,T)\times L^m_\infty(0,\,T)\colon\ F(x,u)=0\}$ where F is the operator

$$F\colon\ \overline{W}^n_{11}(0,\,T)\times L^m_\infty(0,\,T)\to \overline{W}^n_{11}(0,\,T)\times \mathbf{R}^k$$

with $F(x,u)$ given by

$$\left(x(\cdot) - \int_0^{(\cdot)} f(x(s),u(s),s)\ ds,\quad q(x(T))\right).$$

It is easy to see that the operator F has continuous Fréchet derivatives of arbitrary order. For instance, $F'(x,u)$ acting on $(\eta,\xi)\in \overline{W}^n_{11}(0,\,T)\times L^m_\infty(0,\,T)$ is given by

$$\left(\eta(t) - \int_0^t f_x\eta + f_u\xi\ ds,\quad q_x(x(T))\eta(T)\right).$$

All partial derivatives of f are evaluated along a reference input-trajectory pair $(x,\,u)\in\mathcal{A}$. The formulas for higher order derivatives are given by equally straightforward multilinear forms.

We first describe the image of the operator $F'(x_*,\,u_*)$ for a reference input-trajectory pair $(x_*,\,u_*)$. Denote the fundamental matrix of the variational equation by $\Phi(t,s)$, i. e.

$$\frac{\partial}{\partial t}\Phi(t,s) = f_x(x(t),u(t),t)\Phi(t,s),$$
$$\Phi(s,s) = \mathrm{Id}.$$

Furthermore, let $\mathrm{R}\subset\mathbf{R}^n$ denote the reachable subspace of the linearized system

$$\dot{h}(t) = f_x h + f_u v,\quad h(0)=0, \tag{5}$$

at time T. It is well-known that R is a linear subspace of $\mathbf{R}^n$ and that $R = \mathbf{R}^n$ if and only if equation (5) is completely controllable. In general we have that

Lemma 1 *$ImF'(x_*, u_*)$ consists of all pairs $(a, b) \in \overline{W}^n_{11}(0, T) \times \mathbf{R}^k$ such that*

$$b \in q_x(x_*(T)) \left(\int_0^T \Phi(T, s)\dot{a}(s)\, ds + R \right). \tag{6}$$

In particular, $ImF'(x_*, u_*)$ is closed and of finite codimension.

The following characterizations of the nonregularity of the operator F and its codimension are well-known.

Proposition 2 *The codimension of $F'(x_*, u_*)$ is equal to the number of linearly independent solutions to $\dot{\lambda}(t) = -\lambda(t) f_x(x_*(t), u_*(t), t)$ which satisfy $\lambda(t) f_u(x_*(t), u_*(t), t) \equiv$ on [0, T] and for which $\lambda(T)$ is orthogonal to ker $q_x(x_*(T))$.*

Proposition 3 *The operator F is nonregular at $\Gamma = (x_*, u_*)$ if and only if Γ is an abnormal weak extremal which satisfies $H_u(0, \lambda(t), x_*(t), u_*(t), t) \equiv 0$ on [0, T].*

Critical Directions

We describe the set of critical directions along which high-order tangent approximations to the equality constraint $\mathcal{F}$ can be set up. Let $Z = \overline{W}^n_{11}(0, T) \times L^m_\infty(0, T)$ and suppose an admissible process $z_* = (x_*, u_*) \in \mathcal{A}$ and a finite sequence $H_{p-1} = (h_1, \ldots, h_p - 1) \in Z^p - 1$ are given. The following operators allow to formalize high-order approximations to an equality constraint at nonregular points (see, ▶ High-order necessary conditions for optimality for abnormal points). For $k = 1, \ldots, p - 1$, the *directional derivatives* $\nabla^k F(z_*)(H_k)$ of F at z_* along the sequence $H_k = (h_1, \ldots, h_k)$ are given by

$$\sum_{r=1}^{k} \frac{1}{r!} \left(\sum_{j_1 + \cdots + j_r = k} F^{(r)}(z_*)(h_{j_1}, \ldots, h_{j_r}) \right) \tag{7}$$

and we let $G_k[F](z_*; H_{k-1})$ denote the Fréchet-derivatives of the $(k-1)$th directional derivative of F at z_* along H_{k-1}. Thus formally $G_1[F](z_*) = F'(z_*)$ and in general for $k \geq 2$, $G_k = G_k[F](z_*; H_{k-1}) \colon Z \to Y$, $v \to G_k(v)$, is given by

$$G_k(v) = \sum_{r=1}^{k-1} \frac{1}{r!} \times \left(\sum_{j_1 + \cdots + j_r = k-1} F^{(r+1)}(z_*)(h_{j_1}, \ldots, h_{j_r}, v) \right). \tag{8}$$

We also denote by $R_q[F](z_*; H_\ell)$ those terms in the Taylor expansion of $F(z_* + \sum_{i=1}^p \varepsilon^i h_i)$ which are homogeneous of degree $q \geq 2$, but only involve vectors from H_ℓ. The general structure of these remainders is given by

$$\sum_{r=2}^{q} \frac{1}{r!} \left(\sum_{\substack{j_1 + \cdots + j_r = q, \\ 1 \leq j_k \leq \ell, \\ 1 \leq k \leq r}} F^{(r)}(z_*)(h_{j_1}, \ldots, h_{j_r}) \right). \tag{9}$$

Let

$$Y_i = \sum_{k=1}^{i} \operatorname{Im} G_k[F](z_*; H_{k-1}), \quad i = 1, \ldots, p. \tag{10}$$

The following conditions are necessary for the existence of a p-order tangent vector along H_{p-1} [10]:

i) the first $p - 1$ directional derivatives of F along H_{p-1} vanish,

$$\nabla^i F(z_*)(H_i) = 0, \forall i = 1, \ldots, p - 1;$$

ii) the *compatibility conditions*

$$R_{p-1+i}[F](z_*; H_{p-1}) \in Y_i, \quad i = 1, \ldots, p - 1,$$

are satisfied.

In these equations all partial derivatives of f are evaluated along the reference trajectory. These conditions are also sufficient if the operator F is *p-regular* at z_* in direction of the sequence H_{p-1} in the sense of the following definition.

Definition 4 Let $F \colon Z \to Y$ be an operator between Banach spaces. We say the operator F is *p-regular* at z_* in direction of the sequence $H_{p-1} \in Z^{p-1}$ if the following conditions are satisfied:

A1) $F\colon Z \to Y$ is $(2p-1)$-times continuously Fréchet differentiable in a neighborhood of z_*.
A2) The subspaces Y_i, $i = 1, \ldots, p$, are closed.
A3) The map $\mathcal{G}_p = \mathcal{G}_p[F](z_*; H_{p-1})$,

$$\mathcal{G}_p\colon\ Z \to Y_1 \times \frac{Y_2}{Y_1} \times \cdots \times \frac{Y}{Y_{p-1}}$$
$$v \mapsto \mathcal{G}_p(v) = \left(G_1(v), \pi_1 G_2(v), \ldots, \pi_{p-1} G_p(v)\right),$$

where $\pi_i\colon Y_{i+1} \to Y_{i+1}/Y_i$ denotes the canonical projection onto the quotient space, is onto.

In the sense of this definition 1-regularity corresponds to the classical Lyusternik condition while 2-regularity is similar to Avakov's definition [3]. Under these assumptions vectors h_p exist which extend H_{p-1} to p-order tangent vectors to $\mathcal{F}$ at z_* [10,12].

For the critical directions for the objective I we focus on the least degenerate critical case and therefore make the following assumption:

iii) $I'(z_*)$ is not identically zero and $\nabla^i I(z_*)(H_i) = 0$ for $i = 1, \ldots, p-1$.

The assumption that the first $p-1$ directional derivatives vanish is directly tied in with optimality. If there exists a first nonzero directional derivative $\nabla^j I(z_*)(H_j)$ with $j < i$ which is positive, then z_* indeed is a local minimum for any curve $z(\varepsilon) = z_* + \sum_{i=1}^{p} \varepsilon^i h_i + o(\varepsilon^p)$, $\varepsilon > 0$, and none of the directions H_{p-1} is of any use in improving the value. We restrict to $\varepsilon \geq 0$ since we also want to include inequality constraints. On the other hand, if $\nabla^j I(z_*)(H_j) < 0$, then H_j is indeed a direction of decrease and arbitrary high-order extensions of this sequence will give better values. Thus the reference trajectory is not optimal.

We also need to define the critical directions for the inequality constraint $\mathcal{U}$ in the optimal control problem. More generally, we define a p-order feasible set to an inequality constraint in a Banach space.

Definition 5 Let $S \subset Z$ be a subset with nonempty interior. We call v a p-order feasible vector for S at z_* in direction of $H_{p-1} = (h_1, \ldots, h_{p-1}) \in Z^{p-1}$ if there exist an $\varepsilon_0 > 0$ and a neighborhood V of v so that for all $0 < \varepsilon \leq \varepsilon_0$,

$$z_* + \sum_{i=1}^{p-1} \varepsilon^i h_i + \varepsilon^p V \subset S.$$

The collection of all p-order feasible vectors v for S at z_* in direction of the sequence H_{p-1} will be called the *p-order feasible set* to S at z_* in direction of the sequence H_{p-1} and will be denoted by $\mathrm{FS}^{(p)}(S; z_*, H_{p-1})$.

It follows from this definition that $\mathrm{FS}^{(p)}(S; z_*, H_{p-1})$ is open. It is also clear that $\mathrm{FS}^{(p)}(S; z_*, H_{p-1})$ is convex, if S is. Furthermore, if $h_j \in \mathrm{FS}^{(j)}(S; z_*, H_{j-1})$ for some integer $j < p$, then any vector v is allowed as a p-order feasible direction and thus trivially $\mathrm{FS}^{(p)}(S; z_*, H_p - 1) = X$.

For the optimal control problem and $H_{p-1} = ((\eta_1, \xi_1), \ldots, (\eta_{p-1}, \xi_{p-1}))$ let $V_{p-1} = (\xi_1, \ldots, \xi_{p-1}) \in L_\infty^m(0, T)^p$ denote the sequence of controls. Then the critical feasible directions for the convex inequality constraint $\mathcal{U}$ in $L_\infty^m(0, T)$ consist of all H_{p-1} for which

iv) $\mathrm{FS}^{(p)}(\mathcal{U}; u_*, V_{p-1})$ is nonempty.

Definition 6 We call a direction H_{p-1} a *p-regular critical direction* for the extremum problem at z_* if the operator F is p-regular at z_* along H_{p-1} and if conditions (i–iv) are satisfied.

p-Order Local Maximum Principle

Theorem 7 below gives a generalized p-order version of the maximum principle obtained from a dual characterization of the fact that if (x_*, u_*) is optimal, then the p-order tangent cones to the set $\{F = 0\}$, the p-order feasible cone to $\mathcal{U}$ and the p-order cone of decrease for the functional I cannot intersect. Notice that we write covectors like ψ as row vectors. This is consistent with a multiplier interpretation of the adjoint variable. Also we denote partial derivatives by subscripts. For instance, if $\nabla^i f(H_i)$ denotes the ith directional derivative of $f = f(x, u, t)$ with respect to the sequence H_i, then $(\nabla^i f(H_i))_x$ denotes its partial derivative in x. For example, suppose $H_1 = (\eta_1, \xi_1)$. Then

$$\nabla^1 f(H_1) = f_x(x, u, t)\eta_1 + f_u(x, u, t)\xi_1$$

and thus

$$\left(\nabla^1 f(H_1)\right)_x = f_{xx}(x, u, t)\eta_1 + f_{ux}(x, u, t)\xi_1$$

and

$$\left(\nabla^1 f(H_1)\right)_u = f_{xu}(x, u, t)\eta_1 + f_{uu}(x, u, t)\xi_1.$$

Theorem 7 (*p*-order local maximum principle) *Suppose the admissible process (x_*, u_*) is optimal for the optimal control problem (OC). Then for every p-regular critical direction H_{p-1} there exist a number $\nu_0 = \nu_0(H_{p-1}) \geq 0$, vectors $a_i = a(H_{p-1}) \in (\mathbf{R}^k)^*$, $i = 0, \ldots, p-1$, and absolutely continuous functions $\psi(\cdot) = \psi(H_{p-1})(\cdot)$ and $\rho_i(\cdot) = \rho_i(H_{p-1})(\cdot)$, $i = 1, \ldots, p-1$, from $[0, T]$ into $(\mathbf{R}^n)^*$, which satisfy the following conditions along the optimal trajectory $(x_*(t), u_*(t), t)$:*

a) nontriviality condition: ν_0 and the functional λ: $L^m_\infty(0, T) \to \mathbf{R}$, $\xi \longmapsto \lambda(\xi)$, given by

$$\int_0^T \left\langle \nu_0 L_u + \psi f_u + \sum_{i=1}^{p-1} \rho_i \left(\nabla^i f(H_i)\right)_u, \xi \right\rangle dt \quad (11)$$

do not both vanish identically.

b) extended adjoint equation

$$\dot\psi(t) = -\nu_0 L_x - \psi(t) f_x - \sum_{i=1}^{p-1} \rho_i(t) \left(\nabla^i f(H_i)\right)_x \quad (12)$$

with terminal condition

$$\psi(T) = \nu_0 \ell_x(x_*(T)) + a_0 q_x(x_*(T)) + \sum_{i=1}^{p-1} a_i \left(\nabla^i q(x_*(T); H_i)\right)_x. \quad (13)$$

c) orthogonality conditions on the additional multipliers: The functions $\rho_i(\cdot)$, $i = 1, \ldots, p-1$, satisfy

$$\begin{aligned} &\dot\rho_i(t) = -\rho_i(t) f_x, \quad \rho_i(t) f_u \equiv 0, \\ &\rho_i(T) = a_i q_x(x_*(T)) \end{aligned} \quad (14)$$

and for $j = 1, \ldots, i-1$, the following conditions are satisfied for a.e. $t \in [0, T]$:

$$\rho_i(t) \left(\nabla^j f(H_j)\right)_x = 0, \quad (15)$$

$$\rho_i(t) \left(\nabla^j f(H_j)\right)_u = 0, \quad (16)$$

$$a_i \left(\nabla^j q(x_*(1); H_j)\right)_x = 0; \quad (17)$$

d) separation condition: for all vectors $\xi \in FS^{(p)}(\mathcal{U}; u_*, V_{p-1})$ we have that

$$\begin{aligned} 0 \leq\ & \nu_0 R_p[\ell](H_{p-1}) + a_0 R_p[q](H_{p-1}) \\ &+ \sum_{i=1}^{p-1} a_i R_{p+i}[q](H_{p-1}) \\ &+ \int_0^T \left\langle \nu_0 L_u + \psi f_u + \sum_{i=1}^{p-1} \rho_i(t) \left(\nabla^i f(H_i)\right)_u, \xi \right\rangle dt \\ &+ \int_0^T \nu_0 R_p[L](H_{p-1}) + \psi(t) R_p[f](H_{p-1}) \\ &+ \sum_{i=1}^{p-1} \rho_i(t) R_{p+i}[f](H_{p-1})\, dt. \end{aligned} \quad (18)$$

Corollary 8 *The separation condition d) implies the following p-order local minimum condition: along $(x_*(t), u_*(t), t)$ we have for every $u \in U$ and a.e. $t \in [0, T]$:*

$$0 \leq \left\langle \nu_0 L_u + \psi(t) f_u + \sum_{i=1}^{p-1} \rho_i(t) \left(\nabla^i f(H_i)\right)_u,\ u - u_*(t) \right\rangle. \quad (19)$$

In the case of a Lagrangian minimization problem which has no control constraints, or more generally if the control takes values in the interior of the control set, the functional λ vanishes identically. In this case we can normalize $\nu_0 = 1$ and we obtain the following Corollary:

Corollary 9 (*p-order local maximum principle for Lagrangian problems*) *Consider the optimal control problem (OC) without control constraints ($U = \mathbf{R}^m$) and suppose the admissible process (x_*, u_*) is optimal. Then for every p-regular critical direction H_{p-1} there exist vectors $a_i = a(H_{p-1}) \in (\mathbf{R}^k)^*$, $i = 0, \ldots, p-1$, and absolutely continuous functions $\psi(\cdot) = \psi(H_{p-1})(\cdot)$ and $\rho_i(\cdot) = \rho_i(H_{p-1})(\cdot)$, $i = 1, \ldots, p-1$, from $[0, T]$ into $(\mathbf{R}^n)^*$, which satisfy the conditions b)–d) of Theorem 7 along the optimal trajectory $(x_*(t), u_*(t), t)$ for $\nu_0 = 1$. In particular, we thus have*

$$L_u + \psi(t) f_u + \sum_{i=1}^{p-1} \rho_i(t) \left(\nabla^i f(H_i)\right)_u \equiv 0.$$

Example 10 We illustrate Theorem 7 with an example. Consider the problem to minimize the functional $I(x, u)$ given by

$$\int_0^T \left[(x_1 - 1)^2 + x_2^p + (x_3 + 1)^2 - 2\right] dt \tag{20}$$

over all $(x, u) \in \overline{W}^3_{11}(0, T) \times L^2_\infty(0, T)$ subject to the dynamics

$$\dot{x}(t) = \begin{pmatrix} 0 \\ x_1^p \\ \alpha x_2^{p-1} x_3 \end{pmatrix} + \begin{pmatrix} 0 & 1 \\ -1 & 0 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}, \tag{21}$$

initial condition $x(0) = 0$ and terminal constraints $x_1(T) = 0$ and $x_3(T) = 0$. Here p is an integer, $p \geq 2$, and α is an arbitrary real number. For simplicity we have not imposed any control constraints.

It can easily be seen that the reference trajectory $\Gamma = (x_*, u_*) \equiv (0, 0)$ is an abnormal extremal for each problem. In fact, setting $\lambda(t) = (\nu, 0, \nu)$ with $\nu \neq 0$ and $\lambda_0 = 0$ defines an adjoint vector for Γ such that $H_u \equiv 0$. Hence $F'(0, 0)$ is nonregular.

Theorem 7 can be used to eliminate Γ from optimality for any $p \geq 2$. We choose H_{p-1} of the form

$$H_{p-1} = ((\eta_1, \xi_1); (0, 0); \cdots; (0, 0)) \tag{22}$$

with $(\eta_1, \xi_1) \in F'(0, 0)$. With this choice of directions the compatibility conditions ii) simplify considerably and reduce to the first condition only which becomes

$$\int_0^T \left(\eta_1^{[2]}\right)^{p-1} \left(\eta_1^{[3]}\right) ds = 0.$$

Here the superscripts denote the components of the vector η_1. We satisfy this by choosing $\eta_1^{[3]} = -\eta_1^{[1]} \equiv 0$ (i. e., $\xi_1^{[2]} \equiv 0$). Then choosing a nonzero $\eta_1^{[2]}$ with zero boundary conditions defines a nontrivial vector H_{p-1} of the form (22) for which conditions i) and ii) in the definition of p-regular critical directions are satisfied. Furthermore, it is easily seen that the operator F is p-regular in direction of H_{p-1} at Γ. Finally, these directions are also critical for the objective: we have $I'(0, 0)(\eta_1, \xi_1) = 0$ and furthermore

$$\begin{aligned} \nabla^2 I(0, 0)(H_2) &= \frac{1}{2} I''(0, 0)((\eta_1, \xi_1); (\eta_1, \xi_1)) \\ &= \int_0^T \left(\eta_1^{[1]}\right)^2 + \left(\eta_1^{[3]}\right)^2 ds = 0 \end{aligned}$$

provided $p > 2$. Since no other I-derivatives arise in the directional derivatives $\nabla^i I(0, 0)(H_i)$ for $i = 3, \ldots, p - 1$, the direction $H_p - 1 = ((\eta_1, \xi_1); (0, 0); \cdots; (0, 0))$ with $\eta_1^{[1]} = \eta_1^{[3]} \equiv 0$ and a nonzero $\eta_1^{[2]}$ is a nonzero p-regular critical direction for the problem to minimize I subject to $F = 0$ for any $p \geq 2$.

We thus can apply Theorem 7. Since there are no control constraints we can normalize the multipliers so that $\nu_0 = 1$. The additional multipliers ρ_i, $i = 1, \ldots, p - 1$, are associated with elements in the dual spaces of the quotients Y_{i+1}/Y_i (see ▶ High-order necessary conditions for optimality for abnormal points). But here $Y_i = \mathrm{Im}\, F'(0, 0)$ for $i = 1, \ldots, p - 1$, and Y_p is the full space. Thus we have $\rho_i \equiv 0$ for $i = 2, \ldots, p - 1$ and the only nonzero multipliers are ψ and ρ_{p-1} which for simplicity of notation we just call ρ. Now (14) states that ρ is an adjoint multiplier for which the conditions of the local Maximum Principle for an abnormal extremal are satisfied. This multiplier is unique and of the form $\rho(t) = (\nu, 0, \nu)$, but $\nu \in \mathbf{R}$ could be zero. For the extended adjoint equation and minimum condition (19) we need to evaluate the directional derivatives $\nabla^{p-1} f(x, u)(H_i)$. Straightforward, but a bit tedious calculations show that

$$\left(\nabla^{p-1} f(0, 0)(H_i)\right)_x = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \left(\eta_1^{[2]}\right)^{p-1} \end{pmatrix}$$

and

$$\left(\nabla^{p-1} f(0, 0)(H_i)\right)_u \equiv 0.$$

Thus the extended minimum condition reduces to $\psi B \equiv 0$, the minimum condition of the weak maximum principle. Hence also $\psi_2(t) \equiv 0$ and $\psi_1(t) = \psi_3(t)$. But now the extended adjoint equation is given by

$$\dot{\psi}(t) = (2, 0, -2) - \rho \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \left(\eta_1^{[2]}\right)^{p-1} \end{pmatrix}$$

and thus

$$4 = \dot{\psi}_1(t) - \dot{\psi}_3(t) - \nu\left(\eta_1^{[2]}(t)\right)^{p-1} = -\nu\left(\eta_1^{[2]}(t)\right)^{p-1}.$$

But we can certainly choose $\eta_1^{[2]}$ nonconstant to violate this condition. This contradiction proves that Γ cannot be optimal for the problem to minimize I for any $p \geq 2$.

Conclusion

Theorem 7 is based on p-order approximations. If these remain inconclusive, higher order approximations can easily be set up. If the operator F is p-regular in direction of H_{p-1}, then given a p-regular tangent direction, it is possible to set up higher order approximations of arbitrary order. In fact, only a system of p linear equations needs to be solved in every step. These results provide a complete hierarchy of primal constructions of higher-order approximating directions and dual characterizations of empty intersection properties of approximating cones which can be used to give necessary conditions for optimality for increasingly more degenerate structures. For these results see [13].

See also

- ▶ Dynamic Programming: Continuous-time Optimal Control
- ▶ Hamilton–Jacobi–Bellman Equation
- ▶ Pontryagin Maximum Principle

References

1. Agrachev AA, Sarychev AV (1995) On abnormal extremals for Lagrange variational problems. J Math Syst, Estimation and Control 5:127–130
2. Avakov ER (1988) Necessary conditions for a minimum for nonregular problems in Banach spaces. Maximum principle for abnormal problems of optimal control. Trudy Mat Inst Akad Nauk SSSR 185:3–29 (In Russian.)
3. Avakov ER (1989) Necessary extremum conditions for smooth anormal problems with equality-and inequality constraints. J Soviet Math 45. Matematicheskie Zametki 45:3–11
4. Caratheodory C (1935) Variationsrechnung und partielle Differentialgleichungen erster Ordnung. Teubner, Leipzig
5. Dmitruk AV (1998) Quadratic order conditions of a local minimum for abnormal extremals. In: Proc. 2nd World Congress of Nonlinear Analysts, Part 4, Athens 1996. Nonlinear Anal 30:2439–2448
6. Gilbert EG, Bernstein DS (1983) Second-order necessary conditions in optimal control: accessory-problem results without normality conditions. J Optim Th Appl 41:75–106
7. Girsanov IV (1972) Lectures on mathematical theory of extremum problems. Springer, Berlin
8. Ledzewicz U, Schättler H (1997) An extended maximum principle. Nonlinear Anal 29:59–183
9. Ledzewicz U, Schättler H (1998) High order extended maximum principles for optimal control problems with nonregular constraints. In: Hager WW, Pardalos PM (eds) Optimal Control: Theory, Algorithms and Applications. Kluwer, Dordrecht, pp 298–325
10. Ledzewicz U, Schättler H (1998) A high-order generalization of the Lyusternik theorem. Nonlinear Anal 34:793–815
11. Ledzewicz U, Schättler H (1998) A high-order generalization of the Lyusternik theorem and its application to optimal control problems. In: Chen W, Hu S (eds) Dynamical Systems and Differential Equations II. pp 45–59
12. Ledzewicz U, Schättler H (1999) High-order approximations and generalized necessary conditions for optimality. SIAM J Control Optim 37:33–53
13. Ledzewicz U, Schättler H (2000) A high-order generalized local maximum principle. SIAM J Control Optim 38:823–854
14. Milyutin AA (1981) Quadratic conditions of an extremum in smooth problems with a finite-dimensional image. Methods of the Theory of Extremal Problems in Economics. Nauka Moscow, Moscow, pp 138–177 (In Russian.)
15. Pontryagin LS, Boltyanskii VG, Gamkrelidze RV, Mishchenko EF (1962) The mathematical theory of optimal processes. Wiley, New York
16. Stefani G, Zezza PL (1996) Optimality conditions for a constrained control problem. SIAM J Control Optim 34:635–659

High-order Necessary Conditions for Optimality for Abnormal Points

Urszula Ledzewicz[1], Heinz Schättler[2]

[1] Department Math. and Statist., Southern Illinois University at Edwardsville, Edwardsville, USA

[2] Department Systems Sci. and Math., Washington University, St. Louis, USA

MSC2000: 49K27, 46N10, 41A10, 47N10

Article Outline

Keywords

Lyusternik theorem; High-order tangent sets; High-order necessary conditions for optimality; Abnormal processes

We consider the problem of minimizing a functional I: $X \to \mathbf{R}$ in a Banach space X under both *equality* and *inequality constraints*. The inequality constraints are of two types, either described by smooth functionals f: $X \to \mathbf{R}$ as $P = \{x \in X : f(x) \leq 0\}$ or described by closed convex sets C with nonempty interior. The equality constraints are given in operator form as $Q = \{x \in X : F(x) = 0\}$ where F: $X \to Y$ is an operator between Banach spaces. Models of this type are common in optimal control problems.

The standard first order Lagrange multiplier type necessary conditions for optimality at the point x_* state that there exist multipliers $\lambda_0, \ldots, \lambda_m, y^*$ which do not all vanish identically such that the *Euler–Lagrange equation*

$$\lambda_0 I'(x_*) + \sum_{j=1}^{m} \lambda_j f_j'(x_*) + F'^*(x_*)y^* = 0, \qquad (1)$$

is satisfied (see for instance [7,9]). This article addresses the case when the Fréchet-derivative $F'(x_*)$ of the operator defining the equality constraint is not onto, i. e. the regular case. In this case the *classical Lyusternik theorem* [14] does not apply to describe the tangent space to Q and (1) can be satisfied trivially by choosing a nonzero multiplier y^* from the annihilator of Im $F'(x_*)$ while setting all other multipliers zero. This generates so-called *abnormal points* for which the standard necessary conditions for optimality only describe the degeneration of the equality constraint without any relation to optimality. Here we describe an approach to high-order necessary conditions for optimality in these cases which is based a *high-order generalization of the Lyusternik theorem* [12]. By using this theorem one can determine the precise structure of polynomial approximations to Q at x^* when the surjectivity condition on $F'(x_*)$ is not satisfied, but when instead a certain operator G_p which takes into account all derivatives up to and including order p is onto. The order p is chosen as the minimum number for which the operator G_p becomes onto. If G_p is onto, then the precise structure of q-order polynomial approximations to Q at x_* for any $q \geq p$ can be determined. This leads to the notion of high-order tangent cones to the equality constraint Q at points x_* in a nonregular case. Combining these with high-order feasible cones for the inequality constraints and high-order cones of decrease, a generalization of the *Dubovitskii–Milyutin theorem* is formulated. From this theorem generalized necessary conditions for optimality can be deduced which reduce to classical conditions for normal cases, but give new and nontrivial conditions for abnormal cases.

First results of this type have been obtained for quadratic approximations ($p = 2$) in [3,4,5] and [11]. Some of these conditions have been analyzed further also in connection with sufficient conditions for optimality, [1,2]. In [10] also quadratic approximations for problems with inequality constraints are considered. For the regular case when $F'(x_*)$ is onto second order approximating sets were introduced in [6] to derive second order necessary conditions for optimality, while higher order necessary conditions for optimality in this case are given, for instance, in [8] or [15]. These, however, are not the topic of this article.

A High-Order Formulation of the Dubovitskii–Milyutin Theorem

Let X and Y be Banach spaces. Let I: $X \to \mathbf{R}$ be a functional, F: $X \to Y$ an operator, f_j: $X \to \mathbf{R}$, $j = 1, \ldots, m$, functionals and let $C \subset X$ be a closed convex set with nonempty interior. We assume that I, the functionals f_j

and the operator F are sufficiently often continuously Fréchet-differentiable and consider the problem

$$(P)\begin{cases}\min\limits_{x} & I\\ \text{s.t.} & x\in A=\left(\cap_{j=1}^{m}P_j\right)\cap Q\cap C,\\ & P_j=\left\{x\in X\colon\ f_j(x)\le 0\right\}\\ & Q=\left\{x\in X\colon\ F(x)=0\right\}.\end{cases}$$

We define high-order polynomial approximations to the admissible domain A. We denote sequences $(h_1, \ldots, h_k)\in X^k$ by H_k with the subscript giving the length of the sequence.

Definition 1 Let $H_{p-1}=(h_1,\ldots,h_{p-1})\in X^{p-1}$ and set $x(\varepsilon)\doteq x_*+\sum_{i=1}^{p-1}\varepsilon^i h_i$. We call H_{p-1} a $(p-1)$-*order approximating sequence* to a set $S\subseteq X$ at $x_*\in \operatorname{Clos} S$, respectively we call $x{:}\varepsilon\to x(\varepsilon)$, a $(p-1)$-*order approximating curve*, if there exist an $\varepsilon_0>0$ and a function r defined on $[0,\varepsilon_0]$ with values in X, $r\colon [0,\varepsilon_0]\to X$, with the property that

$$x(\varepsilon)+r(\varepsilon)=x_*+\sum_{i=1}^{p-1}\varepsilon^i h_i+r(\varepsilon)\in S \tag{2}$$

and

$$\lim_{\varepsilon\to 0}\frac{\|r(\varepsilon)\|}{\varepsilon^{p-1}}=0. \tag{3}$$

We call a $(p-1)$-order approximating sequence/curve $(p-1)$-*order feasible* if S is an inequality constraint, respectively $(p-1)$-*order tangent* if S is an equality constraint.

Let $x_*\in F$ and assume as given a $(p-1)$-order approximating sequence $H_{p-1}=(h_1,\ldots,h_{p-1})\in X^{p-1}$ with corresponding $(p-1)$-order approximation $x(\varepsilon)\doteq x_*+\sum_{i=1}^{p-1}\varepsilon^i h_i$. It is implicitly assumed that x_* has not been ruled out for optimality. Then we extend the existing $(p-1)$-order approximations to p-order approximations and derive the corresponding necessary conditions for optimality. The following definitions are direct generalizations of standard existing definitions [7].

Definition 2 We call v_0 a *p-order vector of decrease* for a functional $I\colon X\to \mathbf{R}$ at $x_*\in X$ in direction of the sequence $H_{p-1}=(h_1,\ldots,h_{p-1})\in X^{p-1}$ if there exist a neighborhood V of v_0 and a number $\alpha<0$ so that for all $v\in V$ we have

$$I\left(x_*+\sum_{i=1}^{p-1}\varepsilon^i h_i+\varepsilon^p v\right)=I(x(\varepsilon)+\varepsilon^p v)\le I(x_*)+\alpha\varepsilon^p. \tag{4}$$

The collection of all p-order vectors of decrease for I at x_* in direction of the sequence H_{p-1} will be called the *p-order set of decrease* to I at x_* in direction of the sequence H_{p-1} and will be denoted by $\mathrm{DS}^{(p)}(I;x_*,H_{p-1})$.

Definition 3 We call v_0 a *p-order feasible vector* for an inequality constraint P at $x_*\in X$ in direction of H_{p-1} if there exist an $\varepsilon_0>0$ and a neighborhood V of v_0 so that for all $0<\varepsilon\le\varepsilon_0$

$$x_*+\sum_{i=1}^{p-1}\varepsilon^i h_i+\varepsilon^p V=x(\varepsilon)+\varepsilon^p V\subset P. \tag{5}$$

The collection of all p-order feasible vectors v_0 for P at x_* in direction of the sequence H_{p-1} will be called the *p-order feasible set* to P at x_* in direction of the sequence H_{p-1} and will be denoted by $\mathrm{FS}^{(p)}(P;x_*,H_{p-1})$.

Note that by definition the p-order set of decrease to I and the p-order feasible set to P, both at x_* in direction of the sequence H_{p-1}, are open.

Definition 4 We call h_p a *p-order tangent vector* to an equality constraint Q at x_* in direction of the sequence H_{p-1} if $H_p=(h_1,\ldots,h_p)\in X^p$ is a p-order approximating sequence to the set Q at $x_*\in Q$. The collection of all p-order tangent vectors to Q at x_* in direction of the sequence H_{p-1} will be called the *p-order tangent set* to Q at x_* in direction of the sequence H_{p-1} and will be denoted by $\mathrm{TS}^{(p)}(Q;x_*,H_{p-1})$.

These approximating sets can be embedded into cones in the extended state-space $X\times\mathbf{R}$. This has the advantage that many classical results like the Minkowski–Farkas lemma or the annihilator lemma can be directly applied in calculating dual cones (see also [11]). Let us generally refer to p-order sets of decrease, feasible sets and tangent sets as p-order approximating sets and denote them by $\mathrm{AS}^{(p)}(Z;x_*,H_{p-1})$. Then we define the corresponding approximating cones as follows:

Definition 5 Given a p-order approximating set $\mathrm{AS}^{(p)}(Z;x_*,H_{p-1})$ to a set $Z\subset X$ at x_* in direction

of the sequence H_{p-1}, the *p-order approximating cone* to Z at x_* in direction of H_{p-1}, $\mathrm{AC}^{(p)}(Z;x_*, H_{p-1})$, is the cone in $X \times \mathbf{R}$ generated by the vectors $(v, 1) \in \mathrm{AS}^{(p)}(Z;x_*, H_{p-1}) \times \mathbf{R}$.

Thus we talk of the *p-order cone of decrease* for the functional I, *p-order feasible cones* for inequality constraints and *p-order tangent cones* for equality constraints, all at x_* in direction of the sequence H_{p-1}.

Definition 6 Let $C \subseteq Z$ be a cone in a Banach space Z with apex at 0. The *dual* (or *polar*) cone to C consists of all continuous linear functionals $\lambda \in Z^*$ which are nonnegative on C, i. e.

$$C^* = \{\lambda \in Z^* : \ \langle \lambda, v\rangle \geq 0, \ \forall v \in C\}. \tag{6}$$

Then we have

Theorem 7 [11,13] *(p-order Dubovitskii–Milyutin theorem). Suppose the functional I attains a local minimum for problem (P) at $x_* \in A$. Let $H_{p-1} = (h_1, \ldots, h_{p-1}) \in X^{p-1}$ be a $(p-1)$-order approximating sequence such that the p-order cone of decrease for the functional I, the p-order feasible cones for the inequality constraints P_j, $j = 1, \ldots, m$, and C, and the p-order tangent cone to the equality constraint Q, all at x_* in direction of the sequence H_{p-1}, are nonempty and convex. Then there exist continuous linear functionals*

$$\Psi_0 = (\lambda_0, \mu_0) \in \left(\mathrm{DC}^{(p)}(I; x_*, H_{p-1})\right)^*,$$
$$\Psi_j = (\lambda_j, \mu_j) \in \left(\mathrm{FC}^{(p)}(f_j; x_*, H_{p-1})\right)^*,$$

for $j = 1, \ldots, m$,

$$\Omega = (\lambda_{m+1}, \mu_{m+1}) \in \left(\mathrm{FC}^{(p)}(C; x_*, H_{p-1})\right)^*$$

and

$$\Phi = (\lambda_{m+2}, \mu_{m+2}) \in \left(\mathrm{TC}^{(p)}(Q; x_*, H_{p-1})\right)^*,$$

all depending on H_{p-1}, such that

$$\sum_{j=0}^{m+2} \lambda_j \equiv 0, \qquad \sum_{j=0}^{m+2} \mu_j \equiv 0 \tag{7}$$

hold. Furthermore, not all the λ_j, $j = 0, \ldots, m+2$, vanish identically.

High-Order Directional Derivatives

We describe a formalism to calculate higher derivatives [12,13] which will be needed to describe high-order approximating cones. Let $F: X \to Y$ be an operator between Banach spaces which is sufficiently often continuously Fréchet differentiable in a neighborhood of $x_* \in X$ and consider the Taylor expansion of F along a curve

$$\gamma(\varepsilon) = x_* + \sum_{i=1}^{m} \varepsilon^i h_i.$$

We have

$$F(\gamma(\varepsilon)) = F(x_*) + \sum_{i=1}^{m} \varepsilon^i \nabla^i F(x_*)(h_1, \ldots, h_i) + \widetilde{r}(\varepsilon),$$

where $\nabla^i F(x_*)(h_1, \ldots, h_i)$ is given by

$$\sum_{r=1}^{i} \frac{1}{r!} \left(\sum_{j_1 + \cdots + j_r = i} F^{(r)}(x_*)(h_{j_1}, \ldots, h_{j_r}) \right) \tag{8}$$

and $\widetilde{r}(\varepsilon)$ is a function of order $o(\varepsilon^m)$ as $\varepsilon \to 0$. Note that $\nabla^i F(x_*)(h_1, \ldots, h_i)$ simply collects the ε^i-terms in this expansion. These terms, which we call the *ith-order directional derivatives of F along the sequence* $H_i = (h_1, \ldots, h_i)$, $1 \leq i \leq m$, are easily calculated by straightforward Taylor expansions. For example,

$$\nabla^1 F(x_*)(H_1) = F'(x_*)h_1,$$
$$\nabla^2 F(x_*)(H_2) = F'(x_*)h_2 + \frac{1}{2}F''(x_*)(h_1, h_1).$$

The higher-order directional derivative $\nabla^i F(x_*)$ is homogeneous of degree i in the directions in the sense that

$$\nabla^i F(x_*)(\varepsilon h_1, \ldots, \varepsilon^i h_i) = \varepsilon^i \nabla^i F(x_*)(h_1, \ldots, h_i).$$

In particular, no indices j_1 and j_2 with $j_1 + j_2 > i$ can occur together as arguments in any of the terms in $\nabla^i F(x_*)$. Thus all vectors h_j whose index satisfies $2j > i$ appear linearly in $\nabla^i F(x_*)$ and are multiplied by terms which are homogeneous of degree $i - j$. In fact, there exist linear operators $G_k = G_k[F](x_*;H_{k-1})$, $k \in \mathbf{N}$, depending on the derivatives up to order k of F in the point x_* and on the vectors $H_{k-1} = (h_1, \ldots, h_{k-1})$, which describe the contributions of these components. We have $G_1[F](x_*) = F'(x_*)$ and in general

$G_k = G_k[F](x_*;H_{k-1})\colon Z \to Y$, $v \to G_k(v)$, is given by

$$G_k(v) = \sum_{r=1}^{k-1} \frac{1}{r!} \times \left(\sum_{j_1+\cdots+j_r=k-1} F^{(r+1)}(x_*)(h_{j_1},\ldots,h_{j_r},v) \right). \tag{9}$$

These operators $G_k[F](x_*;H_{k-1})$ are the Fréchet-derivatives of the $(k-1)$th directional derivative of F at x_* along H_{k-1}. Note that these terms are homogeneous of degree $k-1$. For simplicity of notation we often suppress the arguments. For example, we write

$$G_1(v) = F'(x_*)v, \quad G_2(v) = F''(x_*)(h_1,v),$$
$$G_3(v) = F''(x_*)(h_2,v) + \frac{1}{2}F'''(x_*)(h_1,h_1,v).$$

Given an order $p \in \mathbf{N}$, it follows that we can separate the linear contributions of the vectors $h_p,\ldots,h_{2p-1}$ in derivatives of orders p through $2p-1$ and for $i=1,\ldots,p$, we have an expression of the form

$$\nabla^{p-1+i}F(x_*)(H_{p-1+i}) = \sum_{k=1}^{i} G_k[F](x_*;H_{k-1})h_{p+i-k} + R_{p-1+i}[F](x_*;H_{p-1}).$$

Here among the terms which are homogeneous of degree $p-1+i$ the sum gives the terms which contain one of the vectors $h_p,\ldots,h_{p-1+i}$, and the remainder R combines all other terms which only include vectors of index $\le p-1$. The general structure of the remainder $R_q[F](z_*;H_\ell)$ for arbitrary $q \ge 2$ and ℓ is given by

$$\sum_{r=2}^{q} \frac{1}{r!} \left(\sum_{\substack{j_1+\cdots+j_r=q,\\ 1\le j_k\le \ell,\\ 1\le k\le r}} F^{(r)}(x_*)(h_{j_1},\ldots,h_{j_r}) \right). \tag{10}$$

Thus $R_q(H_\ell)$ consists of the terms which are homogeneous of degree q, but only involve vectors from H_ℓ. For example, $R_3[F](z_*;H_2)$ is given by

$$F''(z_*)(h_1,h_2) + \frac{1}{6}F^{(3)}(z_*)(h_1,h_1,h_1).$$

Note that the remainders only have contributions from derivatives of at least order two. These operators allow to formalize high-order approximations to an equality constraint at nonregular points [13].

High-Order Tangent Cones

We first describe the set of critical directions along which high-order tangent approximations to the equality constraint Q can be set up. For a given admissible process $z_* \in A$ and a finite sequence $H_{p-1} = (h_1,\ldots,h_{p-1}) \in X^{p-1}$, let

$$Y_i = \sum_{k=1}^{i} \operatorname{Im} G_k[F](x_*;H_{k-1}), \quad i = 1,\ldots,p.$$

It is clear that the first $p-1$ directional derivatives of F along H_{p-1} must vanish,

$$\nabla^i F(z_*)(H_i) = 0, \quad \forall\, i = 1,\ldots,p-1, \tag{11}$$

if H_{p-1} is a $(p-1)$-order tangent direction. But additional compatibility conditions of the form

$$R_{p-1+i}[F](x_*;H_{p-1}) \in Y_i, \quad i = 1,\ldots,p-1, \tag{12}$$

are necessary as well if we want to extend H_{p-1} to a p-order tangent direction $H_p = (H_{p-1};h_p)$. Conditions (11) and (12) are indeed sufficient for the existence of p-order approximations along H_{p-1} under the following regularity condition:

Definition 8 Let $F\colon X \to Y$ be an operator between Banach spaces. We say the operator F is *p-regular* at x_* in direction of the sequence $H_{p-1} \in X^{p-1}$ if the following conditions are satisfied:

A1) $F\colon X \to Y$ is $(2p-1)$-times continuously Fréchet differentiable in a neighborhood of x_*;

A2) the subspaces Y_i, $i=1,\ldots,p$, are closed;

A3) the map $\mathcal{G}_p = \mathcal{G}_p[F](x_*;H_{p-1})$

$$\mathcal{G}_p\colon X \to Y_1 \times \frac{Y_2}{Y_1} \times \cdots \times \frac{Y}{Y_{p-1}},$$
$$v \mapsto \mathcal{G}_p(v) = (G_1(v),\ldots,\pi_{p-1}G_p(v)),$$

where $\pi_i\colon Y_{i+1} \to Y_{i+1}/Y_i$ denotes the canonical projection onto the quotient space, is onto.

In the sense of this Definition, 1-regularity corresponds to the classical Lyusternik condition while 2-regularity is similar to Avakov's definition [5].

Theorem 9 [12] *Let H_{p-1} be a sequence so that $\nabla^i F(x_*)(H_i) = 0$ for $i = 1,\ldots,p-1$, and suppose the operator F is p-regular at x_* in direction of H_{p-1}. Then*

$TS^{(p)}(Q;x_*, H_{p-1})$ is nonempty if and only if for $i = 1, \ldots, p-1$, the compatibility conditions

$$R_{p-1+i}[F](x_*; H_{p-1}) \in Y_i$$

are satisfied. In this case $TS^{(p)}(Q;x_*, H_{p-1})$ is the closed affine subspace of X given by the solutions to the linear equation

$$\mathcal{G}_p[F](x_*; H_{p-1})(v) + \mathcal{R}_{p-1}[F](x_*, H_{p-1}) = 0, \quad (13)$$

where $\mathcal{R}_{p-1}[F](x_*, H_{p-1}) \in Z$ is the point with components

$$\big(R_p[F](x_*; H_{p-1}), \pi_1 R_{p+1}[F](x_*; H_{p-1}), \ldots, \pi_{p-1} R_{2p-1}[F](x_*; H_{p-1})\big).$$

This formulation of the result clearly brings out the geometric structure of the p-order tangent sets as closed affine linear subspaces of X generated by the kernel of $\mathcal{G}_p$, $\ker\mathcal{G}_p$.

Corollary 10 [12] *Let H_{p-1} be a sequence such that the operator F is p-regular at x_* in direction of H_{p-1}. Suppose the first $(p-1)$ directional derivatives $\nabla^i F(x_*)(H_i)$ vanish for $i = 1, \ldots, p-1$, and the compatibility conditions $R_{p-1+i}[F](x_*;H_{p-1}) \in Y_i$ are satisfied for $i = 1, \ldots, p$. Then the p-order tangent cone to $Q = \{x \in X: F(x) = F(x_*)\}$ at x_* in direction of H_{p-1}, $TC^{(p)}(Q;x_*, H_{p-1})$, consists of all solutions $(w, \gamma) \in X \times \mathbf{R}_+$ (i. e. $\gamma > 0$) of the linear equation*

$$\mathcal{G}_p[F](w) + \gamma \mathcal{R}_{p-1}[F](x_*, H_{p-1}) = 0.$$

For applications to optimization problems we need the subspace of continuous linear functionals which annihilate $\mathcal{G}_p$. Since the operator $\mathcal{G}_p$ is onto, it follows by the annihilator lemma or the closed-range theorem [9] that

$$(\ker \mathcal{G}_p)^\perp = \mathrm{Im}(\mathcal{G}_p^*),$$

where $\mathcal{G}_p^*$:

$$Z^* = Y_1^* \times (\frac{Y_2}{Y_1})^* \times \cdots \times (\frac{Y}{Y_{p-1}})^* \to X^*,$$

denotes the adjoint map. Let

$$\tau_i : (\frac{Y_{i+1}}{Y_i})^* \to Y_i^{\perp_{i+1}}$$

denote the canonical isomorphism. Here $\perp_{i+1}$ denotes the annihilator in Y_{i+1}, i. e.

$$Y_i^{\perp_{i+1}} = \{y^* \in Y_{i+1}^* : \ \langle y^*, v\rangle = 0, \ \forall v \in Y_i\}$$

and we formally set $Y_0 = \{0\}$, so that $Y_0^{\perp_1} \cong Y_1^*$. Then we have:

Proposition 11 [11,13] *A functional $\lambda \in X^*$ lies in $(\ker\mathcal{G}_p)^\perp$ if and only if it can be represented in the form*

$$\lambda = \sum_{i=1}^{p} G_i^*[F](x_*; H_{i-1}) y_i^* \quad (14)$$

for some functionals $y_i^* \in Y_{i-1}^{\perp_i}$, $i = 1, \ldots, p$.

Proposition 12 [11,13] *The dual or polar p-order tangent cone consists of all linear functionals $(\lambda, \mu) \in X^* \times \mathbf{R}$ which can be represented in the following form: There exist functionals $y_i^* \in Y_{i-1}^{\perp_i}$, $i = 1, \ldots, p$, and a number $r \geq 0$ such that*

$$\lambda = \sum_{i=1}^{p} G_i^*[F](x_*; H_{i-1}) y_i^*,$$

$$\mu = \sum_{i=1}^{p} \langle y_i^*, R_{p-1+i}[F](x_*; H_{p-1})\rangle + r.$$

High-Order Cones of Decrease

We now consider critical directions for the objective I and determine the p-order sets of decrease of a functional $I: X \to \mathbf{R}$. These results also apply to p-order feasible sets to inequality constraints defined by smooth functionals. We assume as given a $(p-1)$-order sequence H_{p-1} and we calculate the p-order set of decrease of I at x_* along H_{p-1}. Trivial cases arise if there exists a first nonzero directional derivative $\nabla^i I(x_*)(H_i)$ of I with $i \leq p-1$. In this case we have either $DS^{(p)}(I;x_*, H_{p-1}) = \emptyset$ if $\nabla^i I(x_*)(H_i) > 0$ or $DS^{(p)}(I;x_*, H_{p-1}) = X$ if $\nabla^i I(x_*)(H_i) < 0$. In the first case the sequence H_{p-1} cannot be used to exclude optimality of x_* since indeed x_* is a local minimum along the approximating curve generated by H_{p-1}. In the second case h_i is an ith-order direction of decrease along H_{i-1} and thus every vector $v \in X$ is admissible as a pth order component. The only nontrivial case arises

if $\nabla^i I(x_*)(H_i) = 0$ for all i with $i \le p-1$ and if $I'(x_*) \neq 0$.

Proposition 13 [13] *Suppose $I'(x_*) \neq 0$ and for all i with $i \le p-1$ we have $\nabla^i I(x_*)(H_i) = 0$. Then the p-order cone of decrease for the functional I at x_* in direction of H_{p-1}, $DC^{(p)}(I;x_*, H_{p-1})$, consists of all vectors $(w, \gamma) \in X \times \mathbf{R}$ which satisfy*

$$I'(x_*)w + \gamma R_p[I](x_*; H_{p-1}) < 0.$$

Thus $DC^{(p)}(I;x_*, H_{p-1})$ is nonempty, open and convex. The dual or polar cone to $DC^{(p)}(I;x_*, H_{p-1})$ can easily be calculated using the Minkowski–Farkas lemma [7].

High-Order Feasible Cones to Inequality Constraints Given by Smooth Functionals

In this section we give the form of the p-order feasible cones, $FC^{(p)}(P;x_*, H_{p-1})$, for inequality constraints P described by smooth functionals,

$$P = \{x \in X:\ f(x) \le 0\}.$$

Similar like for sets of decrease, if there exists a first index $i \le p-1$ such that $\nabla^i f(x_*)(H_i) \neq 0$, then the constraint will either be satisfied for any p-order vector $v \in X$ if $\nabla^i f(x_*)(H_i) < 0$ or it will be violated if $\nabla^i f(x_*)(H_i) > 0$. This leads to the definition of p-order active constraints.

Definition 14 The inequality constraint P is said to be *p-order active* along the sequence H_{p-1} if for all i, $i = 1, \ldots, p-1$, we have $\nabla^i f(x_*)(H_i) = 0$.

Only p-order active constraints enter the necessary conditions for optimality derived via p-order approximations along an admissible sequence H_{p-1}; p-order inactive constraints generate zero multipliers since $DS^{(p)}(P;x_*, H_{p-1}) = X$ (p-order complementary slackness conditions) and can be ignored for high-order approximations.

Proposition 15 *If the constraint $P = \{x \in X: f(x) \le 0\}$ is p-order active along the sequence H_{p-1}, then the p-order feasible cone, $FC^{(p)}(P;x_*, H_{p-1})$, consists of all vectors $(w, \gamma) \in X \times \mathbf{R}_+$ which satisfy*

$$f'(x_*)w + \gamma R_p[f](x_*; H_{p-1}) < 0.$$

Hence, if $f'(x_*) \neq 0$, then $FC^{(p)}(P;x_*, H_{p-1})$ is nonempty, open and convex.

High-Order Feasible Cones to Closed Convex Inequality Constraints

Let $C \subset X$ be a closed convex set with nonempty interior. Again we assume that H_{p-1} is a $(p-1)$-order feasible sequence. Note that it follows from Definition 3 that $FS^{(p)}(C;x_*, H_{p-1})$ is open (since any vector in the neighborhood V of v also lies in $FS^{(p)}(C;x_*, H_{p-1})$). It is also clear that $FS^{(p)}(C;x_*, H_{p-1})$ is convex, since C is. Thus $FC^{(p)}(C;x_*, H_{p-1})$ is an open, convex cone. Furthermore, if there exists an integer $j <$ p so that $h_j \in FS^{(j)}(C;x_*, H_{j-1})$, then any vector v is allowed as a p-order feasible direction and thus trivially $FS^{(p)}(C;x_*, H_{p-1}) = X$, i. e. the convex constraint $x \in C$ is not p-order active. In this case the necessary conditions for optimality along H_{p-1} are exactly the same as without C.

The dual or polar cone $FC^{(p)}(C;x_*, H_{p-1})^*$ can be identified with all supporting hyperplanes to $FS^{(p)}(C;x_*, H_{p-1})$ at x_*. More precisely, it consists of all linear functionals $(\lambda, \mu) \in X^* \times \mathbf{R}$ which satisfy

$$\langle \lambda, v \rangle + \mu \ge 0, \quad \forall v \in FS^{(p)}(C; x_*, H_{p-1}).$$

Corollary 16 [13] *Let $C \subset X$ be a closed convex set with nonempty interior and suppose the p-order feasible set $FS^{(p)}(C;x_*, H_{p-1})$ is nonempty. If $(\lambda, \mu) \in FC^{(p)}(C;x_*, H_{p-1})^*$, then λ is a supporting hyperplane to C at x_*.*

Generalized Necessary Conditions for Optimality

We now give generalized necessary conditions for optimality for problem (P) based on general p-order approximations. We assume as given a sequence $H_{p-1} = (h_1, \ldots, h_{p-1}) \in X^{p-1}$ with the following properties:

P1) The first $p-1$ directional derivatives of F along H_{p-1} vanish,

$$\nabla^i F(x_*)(H_i) = 0, \quad \forall i = 1, \ldots, p-1,$$

the compatibility conditions

$$R_{p-1+i}[F](x_*; H_{p-1}) \in Y_i$$

are satisfied for $i = 1, \ldots, p-1$, and the operator F is p-regular at x_* in direction of the sequence H_{p-1}.

P2) Either the first nonvanishing derivative $\nabla^i I(x_*)(H_i)$ is negative or $\nabla^i I(x_*)(H_i) = 0$ for $i = 1, \ldots, p-1$.

P3) If the jth constraint is not p-order active, then the first nonzero derivative $\nabla^i f(x_*)(H_i)$ is negative.

P4) $FS^{(p)}(C; x_*, H_{p-1})$ is nonempty.

These conditions guarantee respectively that the corresponding p-order approximating cones to the constraints or the functional I are nonempty and convex. The next theorem generalizes the classical first order necessary conditions for optimality for a mathematical programming problem with convex inequality constraints [7, Thm. 11.4].

Theorem 17 *If x_* is optimal for problem (P), then given any sequence $H_{p-1} = (h_1, \ldots, h_{p-1}) \in X^{p-1}$ for which conditions P1)–P4) are satisfied, there exist Lagrange multipliers $\nu_i \geq 0$, $i = 0, \ldots, m$, functionals $y_i^* \in Y_{i-1}^{\perp_i}$, $i = 1, \ldots, p$, and a supporting hyperplane $\langle \lambda, \nu \rangle + \mu \geq 0$ for all $\nu \in FS^{(p)}(C; x_*, H_{p-1})$, all depending on the sequence H_{p-1}, such that the multipliers ν_i, $i = 0, \ldots, m$, and λ do not all vanish, and*

$$\lambda \equiv \nu_0 I'(x_*) + \sum_{j=1}^{m} \nu_j f_j'(x_*) + \sum_{i=1}^{p} G_i^* y_i^*, \tag{15}$$

$$\mu \leq \nu_0 R_p[I](x_*; H_{p-1}) + \sum_{j=1}^{m} \nu_j R_p[f_j](x_*; H_{p-1}) + \sum_{i=1}^{p} \left\langle y_i^*, R_{p-1+i}[F](H_{p-1}) \right\rangle. \tag{16}$$

Furthermore, the following p-order complementary slackness conditions hold:

- *$\nu_0 = 0$ if $DS^{(p)}(I; x_*, H_{p-1}) = X$;*
- *$\nu_j = 0$ if $FS^{(p)}(P_j; x_*, H_{p-1}) = X$;*
- *$\lambda = 0$ if $FS^{(p)}(C; x_*, H_{p-1}) = X$.*

Remark 18 This theorem gives the formulation for the case which is *nondegenerate* in the sense that the operator $\mathcal{G}_p$ is onto and it is this condition which implies the nontriviality of the multipliers ν_j, $j = 0, \ldots, m$, and λ. If $\mathcal{G}_p$ is not onto, but $\mathrm{Im}\mathcal{G}_p$ is closed, while all the other conditions remain in effect, then a degenerate version of this theorem can easily be obtained by choosing a nontrivial multiplier $\widetilde{y}^* \in (\mathrm{Im}\,\mathcal{G}_p)^{\perp}$. This then gives rise to nontrivial multipliers $y_i^* \in Y_{i-1}^{\perp_i}$ which have the property that $\sum_{i=1}^{p} G_i^* y_i^* \equiv 0$. Thus (15) still holds if we set $\nu_j = 0$, for $j = 0, \ldots, m$, and $\lambda = 0$. Thus the difference is that it can only be asserted that not all of the multipliers $\nu_j, j = 0, \ldots, m$, $y_i^* \in Y_{i-1}^{\perp_i}$, $i = 1, \ldots, p$, and λ do vanish.

See also

► Kuhn–Tucker Optimality Conditions

References

1. Arutyunov AV (1991) Higher-order conditions in anormal extremal problems with constraints of equality type. Soviet Math Dokl 42(3):799–804
2. Arutyunov AV (1996) Optimality conditions in abnormal extremal problems. System Control Lett 27:279–284
3. Avakov ER (1985) Extremum conditions for smooth problems with equality-type constraints. USSR Comput Math Math Phys 25(3):24–32. (Zh Vychisl Mat Fiz 25(5))
4. Avakov ER (1988) Necessary conditions for a minimum for nonregular problems in Banach spaces. Maximum principle for abnormal problems in optimal control. Trudy Mat Inst Akad Nauk SSSR 185:3–29; 680–693 (In Russian.)
5. Avakov ER (1989) Necessary extremum conditions for smooth anormal problems with equality-and inequality constraints. J Soviet Math 45:3–11. (Matematicheskie Zametki 45)
6. Ben-Tal A, Zowe J (1982) A unified theory of first and second order conditions for extremum problems in topological vector spaces. Math Program Stud 19:39–76
7. Girsanov IV (1972) Lectures on mathematical theory of extremum problems. Springer, Berlin
8. Hoffmann KH, Kornstaedt HJ (1978) Higher-order necessary conditions in abstract mathematical programming. J Optim Th Appl (JOTA) 26:533–568
9. Ioffe AD, Tikhomirov VM (1979) Theory of extremal problems. North-Holland, Amsterdam
10. Izmailov AF (1994) Optimality conditions for degenerate extremum problems with inequality-type constraints. Comput Math Math Phys 34:723–736
11. Ledzewicz U, Schättler H (1995) Second-order conditions for extremum problems with nonregular equality constraints. J Optim Th Appl (JOTA) 86:113–144
12. Ledzewicz U, Schättler H (1998) A high-order generalization of the Lyusternik theorem. Nonlinear Anal 34:793–815
13. Ledzewicz U, Schättler H (1999) High-order approximations and generalized necessary conditions for optimality. SIAM J Control Optim 37:33–53
14. Lyusternik LA (1934) Conditional extrema of functionals. Math USSR Sb 31:390–401
15. Tretyakov AA (1984) Necessary and sufficient conditions for optimality of p-th order. USSR Comput Math Math Phys 24(1):123–127

Hilbert's Thirteenth Problem

VICTOR KOROTKICH
Central Queensland University, Mackay, Australia

MSC2000: 01A60, 03B30, 54C70, 68Q17

Article Outline

Keywords

Superpositions of functions; Algebraic equations; E-entropy; Information

The formulation of Hilbert's thirteenth problem [8] reads: 'impossibility of solving the general equation of degree 7 by means of any continuous functions depending only on two variables' [21].

On this basis, D. Hilbert proposed that the complexity of functions is specified essentially by the number of variables. However, as turned out later, this proposal being valid for analytic functions is not true in the general case. In particular, complexity of r times continuously differentiable functions of n variables depends not on the number of variables n but on the ratio n/r.

It is known that the equation of third degree can be reduced by translation to

$$X^3 + pX + q = 0,$$

which has the solution (S. del Ferro, 16th century)

$$X = \left[-\frac{q}{2} + \sqrt{\frac{4p^3 + 27q^2}{4(27)}}\right]^{1/3} + \left[-\frac{q}{2} - \sqrt{\frac{4p^3 + 27q^2}{4(27)}}\right]^{1/3}.$$

The equation of fourth degree can be solved by superposition of addition, multiplication, square roots, cube roots and fourth roots.

To try to solve *algebraic equations* of higher degree (a vain hope according to N.H. Abel and E. Galois), the idea of W. Tschirnhausen in 1683 [24] was to adjoin a new equation, i. e., to

$$P(X) = 0$$

one adjoins

$$Y = Q(X),$$

where Q is a polynomial of degree strictly less than that of P, chosen expediently. In this way one can show that the roots of an equation of degree 5 can be expressed via the usual arithmetic operations in terms of radicals and of the solution $\phi(x)$ of the quintic equation

$$X^5 + xX + 1 = 0$$

depending on the parameter x. Similarly for the equation of degree 6, the roots are expressible in the same way if we include also a function $\theta(x, y)$, a solution of a 6th-degree equation depending on two parameters x and y.

For degree 7 we would have to include also a function $\sigma(x, y, z)$, solution of the equation

$$X^7 + xX^3 + yX^2 + zX + 1 = 0.$$

Hence the natural question: Can $\sigma(x, y, z)$ be expressed by superposition of algebraic functions of two variables [10]?

A great number of papers are devoted to the representability of functions as *superpositions of functions* depending on a smaller number of variables and satisfying certain additional conditions such as algebraicity, analyticity and smoothness. Hilbert was aware of the fact that superpositions of discontinuous functions represent all functions of a larger number of variables. He also knew about the existence of analytic functions of three variables that cannot be represented by any finite superpositions of analytic functions of two variables [8].

In the statement of his 13th problem, Hilbert proceeded from a result of Tschirnhausen [24], according to which a root of an algebraic equation of degree $n > 5$, i. e., a function $f(x_1, \ldots, x_n)$ determined by an equation

$$f^n + x_1 f^{n-1} + \cdots + x_n = 0, \tag{1}$$

can be expressed as a superposition of algebraic functions of $n-4$ variables [21]. Hilbert assumed that the

number $n-4$ cannot be reduced for $n = 6, 7, 8$ and also proved that in order to solve an equation of degree $n = 9$ it suffices to have functions of $n-5$ variables [9]. A. Wiman [26] extended the latter result to $n > 9$, while N. Chebotarev [6] reduced the number of variables involved in the representation of functions to $n-6$ for $n \geq 21$ and to $n-7$ for $n \geq 121$.

Chebotarev was the first to attempt to find topological obstructions to the representability of algebraic functions as superpositions of algebraic functions, but his proofs were not convincing [5,17]. Using topological notions related to the behavior of a many-valued algebraic function on and near a branching manifold, it is proved that algebraic functions cannot be represented by complete superpositions of integral algebraic functions. Completeness means that the represented function must involve all the branches of the many-valued functions and not only one of them as, for example, in the formulas expressing solutions to equations of the 3rd and the 4th degree [21].

Certain topological obstructions to the representation by a complete superpositions of algebraic functions were constructed in this way [2]. V. Lin [15] established the following, most complete, result: In any neighborhood of the origin for $n \geq 3$ the root $f(x_1, \ldots, x_n)$ of equation (1) is not a complete superposition of entire algebroid functions of fewer than $n-1$ variables and single-valued holomorphic functions of an arbitrary number of variables. Thus, from the standpoint of complete superpositions of entire algebraic functions, even fourth-degree equations cannot be solved without using functions of three variables [21].

Hilbert had had another motivation for his thirteenth problem: *nomography*, the method of solving equations by drawing a one-parameter family of curves. This problem, arising in the methods of computation of Hilbert's time, inspired the development of Kolmogorov's notion of *ε-entropy* [20]. Applications of ε-entropy have its crucial role in theories of approximation now used in computer science [22].

In *Kolmogorov ε-entropy*, a natural characteristic of a function class F is

$$H_\varepsilon(F) = \log_2 N_\varepsilon(F),$$

where $N_\varepsilon(F)$ is the minimum number of points in an ε-net in F. Broadly speaking, the ε-entropy of a function class F is the amount of *information* needed to specify with accuracy ε a function of the class F. A main problem in ε-entropy is estimates for the rate of growth of $H_\varepsilon(F)$ as $\varepsilon \to 0$ for Lipschitz functions, classes of analytic functions and functions possessing a given number of derivatives. A.N. Kolmogorov showed that the ε-entropy of r times continuously differentiable functions of n variables grows as $\varepsilon^{-n/r}$ [20].

Since a digital computer can store only a finite set of numbers, functions must be replaced by such finite sets. Therefore, studies in ε-entropy are important for the correct estimation of the possibilities of computational methods for approximately representing functions, their implementation on computers and their storage in the computer memory.

Also ε-entropy has many other applications [23]. An ε-net of Lipschitz functions of n variables is constructed to design global optimization algorithms. This ε-net is based on the Kolmogorov's minimal ε-net of one-dimensional Lipschitz functions and is encoded in terms of monotone functions of k-valued logic. This construction gives a representation of an n-dimensional global optimization problem by a minimal number of one-dimensional ones without loss of information [13].

Let us briefly recall the history of the solution of the Hilbert's thirteenth problem by Kolmogorov and V. Arnol'd. Hilbert's problem was first solved on the basis of ideas by using technique developed by A. Kronrod [14]. In this way Kolmogorov proved that any continuous function of $n \geq 4$ variables can be represented as a superposition of continuous functions of three variables [11]. For an arbitrary function of four variables the representation has the form

$$f(x_1, x_2, x_3, x_4) = \sum_{r=1}^{4} h^r[x_4, g_1^r(x_1, x_2, x_3), g_2^r(x_1, x_2, x_3)].$$

The question whether an arbitrary continuous function of three variables can be represented as a superposition of continuous functions of two variables remained open. The method reduced the representability of functions of three variables as superpositions of functions of two variables to a representability problem for functions defined on universal trees of three-dimensional space [21].

Contrary to the expectations of Hilbert and of his contemporary mathematicians, in 1957 Arnol'd [1], who was a student of Kolmogorov, solved the latter problem and gave the final solution to Hilbert's thirteenth problem in the form of a theorem asserting that any continuous function of $n \geq 3$ variables can be represented as a superposition of functions of two variables [21].

A few weeks later Kolmogorov showed that any continuous function f of n variables can be represented as a superposition

$$f(x_1, \ldots, x_n) = \sum_{q=1}^{2n+1} \chi_q \left[\sum_{p=1}^{n} \phi^{pq}(x_p) \right] \qquad (2)$$

of continuous functions of one variable and the operation of addition [12]. In Kolmogorov's representation (2) the inner functions ϕ^{pq} are fixed and only the outer functions χ_q depend on the represented function f.

The results of [11] do not follow from the theorem presented in [12] in their exact statements, but their essence (in the sense of the possibility of representing functions of several variables by means of superpositions of functions of a smaller number of variables and their approximation by superpositions of a fixed form involving polynomials in one variable and addition) is obviously contained in it [12]. The method for proving the theorem is more elementary than that in [1,11] and reduces to direct constructions and calculations. In Kolmogorov's opinion, the proof of the theorem was his most technically difficult achievement [21].

Thorough proofs of Kolmogorov's theorem and the lemmas of his paper [12] were published in [16,18,20] and others. G. Lorenz [16] noted that the outer functions χ_q can be replaced by a single function χ. D. Sprecher [18] reduced all the inner functions to translations and extensions of a single function ψ with the property that there exits $\varepsilon > 0$ and $\lambda > 0$ such that any continuous function of n variables can be represented as

$$f(x_1, \ldots, x_n) = \sum_{q=1}^{2n+1} \chi[\lambda^p \psi(x_p + \varepsilon q) + q)]. \qquad (3)$$

B. Fridman [7] proved that the inner functions ϕ^{pq} in (2) can be chosen so that they satisfy a Lipschitz condition. Sprecher [19] extended this result to the representation (3) (the function ψ can be chosen to satisfy a Lipschitz condition).

It follows from Kolmogorov's representation (2) and Bari's representation [3] of any continuous function of one variable as a sum of three superpositions of absolutely continuous functions $\sum f_k \circ gk$ that all continuous functions of any number of variables can be represented by means of superpositions of absolutely continuous functions of one variable and the operation of addition [21].

In the opposite direction are the results of A. Vitushkin [25] and L. Bassalygo [4]. When we deal with superpositions of formal series or analytic functions it can be shown that, for example, almost every entire function has at an arbitrary point of C^3 a germ which is not expressible by superposition of series in two variables. So there are many more entire functions of three variables than of two [10]. The result of Vitushkin is that there exist r times continuously differentiable functions of n variables that cannot be expressed in terms of finite superpositions of $s \geq 1$ times continuously differentiable functions of $k < n$ variables if $n/r > ks$ [25], representability depends on n/r. Bassalygo proved that for any three functions ψ_k continuous on a square there exists a continuous function f which cannot be represented as $\sum \chi_k \circ \psi_k$ for any continuous χ_k [4].

See also

► History of Optimization

References

1. Arnold V (1957) On the representation of continuous functions of three variables as superpositions of continuous functions of two variables. Dokl Akad Nauk SSSR 114(4):679–681(in Russian.)
2. Arnold V (1970) On cohomology classes of algebraic functions invariant under a Tschirnhausen transformation. Funkts Anal i Prilozhen 4(1):84–85(in Russian.)
3. Bari N (1930) Memoire sur la representation finie des fonctions continues. Math Anal 103:185–248
4. Bassalygo L (1966) On the representation of continuous functions of two variables by continuous functions of one variable. Vestn MGU Ser Mat-Mekh 21:58–63(in Russian).
5. Chebotarev N (1943) The resolvent problem and critical manifolds. Izv Akad Nauk SSSR Ser Mat 7:123–146(in Russian.)
6. Chebotarev N (1954) On the resolvent problem. Uchen Zap Kazan Univ 114(2):189–193(in Russian.)

7. Fridman B (1967) Increasing smoothness of functions in Kolmogorov's theorem on superpositions. Dokl Akad Nauk SSSR 177:1019–1022(in Russian.)
8. Hilbert D (1902) Sur les problemes futurs des mathematiques. Proc. Second Internat. Congress of Mathematicians, Gauthier-Villars, 58–114
9. Hilbert D (1927) Ueber die Gleichung neunten Grades. Math Ann 97:243–250
10. Kantor J (1996) Hilbert's problems and their sequels. Math Intelligencer 18(1):21–30
11. Kolmogorov A (1956) On the representation of continuous functions of several variables as superpositions of continuous functions of a smaller number of variables. Dokl Akad Nauk SSSR 108(2):179–182 (in Russian.)
12. Kolmogorov A (1957) On the representation of continuous functions of several variables as superpositions of continuous functions of one variable and addition. Dokl Akad Nauk SSSR 114(5):953–956 (in Russian.)
13. Korotkich V (1990) Multilevel dichotomy algorithm in global optimization. In: Sebastian H, Tammer K (eds) System Modelling and Optimization. Springer, Berlin, pp 161–169
14. Kronrod A (1950) Uspekhi Mat Nauk 5(1):24–134(in Russian.)
15. Lin V (1976) Superpositions of algebraic functions. Funkts Anal i Prilozhen 10(1):37–45(in Russian.)
16. Lorenz G (1962) Metric entropy, width and superpositions of functions. Amer Math Monthly 69:469–485
17. Morozov V (1954) On some questions in the resolvent problem. Uchen Zap Kazan Univ 114(2):173–187(in Russian.)
18. Sprecher D (1965) On the structure of continuous functions of several variables. Trans Amer Math Soc 115:340–355
19. Sprecher D (1972) An improvement in the superposition theorem of Kolmogorov. J Math Anal Appl 38:208–213
20. Tikhomirov V (1963) A.N. Kolmogorov's work on ε-entropy of function classes and superpositions of functions. Uspekhi Mat Nauk 18(5):55–92(in Russian.)
21. Tikhomirov V (ed) (1991) Selected works of A.N. Kolmogorov: Mathematics and mechanics, vol 1. Kluwer, Dordrecht
22. Traub J, Wasilkowski G, Wozniakowski H (1988) Information-based complexity. Acad. Press, New York
23. Traub J, Wozniakowski H (1980) Theory of optimal algorithms. Acad. Press, New York
24. Tschirnhausen W (1683) Methodus auferendi omnes terminos intermedios ex data equatione. Acta Eruditorum
25. Vitushkin A (1955) On multidimensional variations. Gostekhteorizdar
26. Wiman A (1927) Ueber die Anwendung der Tschirnhausen Transformation auf die Reduktion algebraischer Gleichungen. Nova Acta R Soc Sci Uppsaliensis extraordin. edit.: 3–8

History of Optimization

Ding-Zhu Du[1], Panos M. Pardalos[2], Weili Wu[3]

[1] Department Computer Sci. and Engineering, University Minnesota, Minneapolis, USA

[2] Center for Applied Optim. Department Industrial and Systems Engineering, University Florida, Gainesville, USA

[3] Department Computer Sci. and Engineering, University Minnesota, Minneapolis, USA

MSC2000: 01A99

Article Outline

Keywords

History; Optimization

Did you ever watch how a spider catches a fly or a mosquito? Usually, a spider hides at the edge of its net. When a fly or a mosquito hits the net, the spider will pick up each line in the net to choose the tense one and then goes rapidly along the line to its prey. Why does the spider chooses the tense line? Some biologists explain that the line gives the shortest path from the spider to its prey.

Did you heard the following story about a wise general? He had a duty to capture a town behind a mountain. When he and his soldiers reached the top of the mountain, he found that his enemy had already approached the town very closely from another way. His dilemma was how to get in the town before the enemy arrive. It was a challenging problem for the general. The general solved the problem by asking each soldier to roll down the mountain in a blanket. Why is this faster? Physicists tell us that a free ball rolling down a mountain always chooses the most rapid way.

Do you know the tale of a horse match of Tian Gi? It is a story set in BC time. Tian Gi was a general in one of several small counties of China, called Qi. The King of Qi knew that Tian Gi had several good horses and ordered Tian Gi to have a horse match with him. The match consisted of three rounds. In each round, each

side chose a horse to compete with the other side. Tian Gi knew that his best horse could not compete with the best one of the King, his second best horse could not compete with the second best one of King, and his third best horse could not compete with the third best one of the King. Therefore, he did not use his best horse against the best horse of the King. Instead, he put his third best horse in the first round against the best one of the King, his best horse in the second round against the second best one of the King, and his second best horse in the third round against the third best one of the King. The final result was that although he lost the first round of the match, he won the last two rounds. Tian Gi's strategy was the best to win this match. Today, economists tell us that many economic systems and social systems can be modeled into games. Each contestant in the game tries to maximize certain benefits.

Optimality is a fundamental principle, establishing natural lows, ruling biologic behaviors, and conducting social activities. Therefore, optimization started from the earliest stages of human civilization. Of course, before mathematics was well established, optimization could be done only by simulation. One may find many wise men's stories in the human history about it. For example, to find the best way to get out of a mountain, someone followed a stream, and to find the best way to get out from a desert, someone set an old horse free and followed the horse's trace.

In the 19th century or even today, simulation is still used for optimizing something. For example, to find a shortest path on a network, one may make a net with rope in a proportional size and pull the net tightly between two destinations. The tense rope shows the shortest path. To find an optimal location of a school for three villages, one may drill three holes on a table and put a piece of rope in each hole. Then tie three rope-ends above the table together and hang a one-kg-weight on each rope-end under the table. When this mechanical system is balanced, the knot of the three rope-pieces points out the location of the school.

The history of optimization in mathematics can be divided into three periods.

In the first period, one did not know any general method to find a maximum/minimum point of a function. Only special techniques were found to maximize/minimize some special functions. A typical function is the quadratic function of one variable

$$y = ax^2 + bx + c.$$

The study of quadratic functions was closely related to the study of constantly-accelerating movement. What is the highest point that a stone is thrown out with certain initial speed and certain angle? What is the farthest point where a stone thrown with certain initial speed can reach when throwing angle varies? These were questions considered by some physicists and generals. In fact, the stone-throwing machine was an important weapon in military.

Today (as of 2000), computing maximum/minimum points of a quadratic function is still an important technique of optimization, existing in elementary mathematics books. The technique had been also extended to other functions such as

$$y = \frac{x^2 + x + 1}{x^2 + 2x + 3}.$$

Actually, multiplying both sides by $x^2+ 2x+3$ and simplifying, we obtain

$$(y-1)x^2 + (2y-1)x + (3y-1) = 0.$$

Since x is a real number, we must have

$$(2y-1)^2 - 4(y-1)(3y-1) \geq 0.$$

Therefore,

$$-8y^2 + 12y - 3 \geq 0,$$

that is,

$$2(3-\sqrt{3}) \leq y \leq 2(3+\sqrt{3}).$$

It is interesting to note that with this technique we obtained the global maximum and minimum of y.

A new period started in 1646 by P. de Fermat. He proposed, in his paper [5], a general approach to compute local maxima/minima points of a differentiable function, that is, setting the derivative of the function to be zero. Today, this approach is still included in almost all textbooks of calculus as an application of differentiation. In this period, optimization existed scattered and disorderly in mathematics. Because optimization had not become an important branch of applied mathematics, some mathematicians did not pay so much attention to results on optimization and some contributions

were even not put in any publication. This left many mysteries in the history of optimization.

For example, who is the first person who proposed the Steiner tree? It was one such mystery. To obtain a clear view, let us explain it in a little detail.

In the same paper mentioned above, Fermat also studied a problem of finding a point to minimize the total distance from it to three given points in the Euclidean plane. Suppose three given points are (x_1, y_1), (x_2, y_2), and (x_3, y_3). Then the total distance from a point (x, y) to these three points is

$$f(x, y) = \sum_{i=1}^{3} \sqrt{(x - x_i)^2 + (y - y_i)^2}.$$

By Fermat's general method, the minimum point of $f(x, y)$ must satisfy the following equations

$$\frac{\partial f}{\partial x} = \sum_{i=1}^{3} \frac{x - x_i}{\sqrt{(x - x_i)^2 + (y - y_i)^2}} = 0,$$

$$\frac{\partial f}{\partial y} = \sum_{i=1}^{3} \frac{y - y_i}{\sqrt{(x - x_i)^2 + (y - y_i)^2}} = 0.$$

However, obtaining x and y from this system of equations seems hopeless. Therefore, Fermat mentioned this problem again in a letter to A. Mersenne that it would be nice if a clear solution could be obtained for this problem.

E. Torricelli, a student of G. Galilei, obtained a clever solution with a geometric method. He showed that if three given points form a triangle without an angle of at least 120°, then the solution is a point at which three segments from it to three given points produce three angles of 120°. Otherwise, the solution is the given point at which the triangle formed by the three given points has an angle of at least 120°.This result can also be proved by the mechanic system described at the beginning of this article. In the first case, the knot of the three rope-pieces stays not at any given point and hence the balance condition of the three forces of equal magnitude yields the condition on the angles. In the second case, the knot falls in one of the three holes, and the condition on the angle guarantees that the knot would not move away from the hole.

Fermat's problem was extensively studied later and was generalized to four points by J.Fr. Fagnano in 1775 and to n points by P. Tedenat and S. L'Huiller in 1810. Fagnano pointed out that it is very easy to find the solution of Fermat's problem for four points. When four given points form a convex quadrilateral, the solution of Fermat's problem is the intersection of two diagonals, i. e., the intersection of two diagonals minimizes the total distance from one point to four given points. Otherwise, there must be one of the given points lying inside the triangle formed by the other three given points; this given point is the solution.

On March 19, 1836, H.C. Schumacher wrote a letter to C.F. Gauss. In his letter, he mentioned a paradox about Fermat's problem: Consider a convex quadrilateral $ABCD$. It has been known that the solution of Fermat's problem for four points A, B, C, and D is the intersection E of diagonals AC and BD. Suppose extending DA and CB can obtain an intersection F. Now, moving A and B to F. Then E will also be moved to F. However, when the angle at F is less than 120°, the point F cannot be the solution of Fermat's problem for three given points F, D, and C.What happens?

On March 21, 1836, Gauss wrote a letter to Schumacher in which he explained that the mistake of Schumacher's paradox occurs at the place where Fermat's problem for four points A, B, C, and D is changed to Fermat's problem for three points F, C, and D. When A and B are identical to F, the total distance from E to four points A, B, C, and D equals $2EF + EC + ED$, not $EF + EC + ED$. Thus, the point E may not be the solution of Fermat's problem for F, C, and D. More importantly, Gauss proposed a new problem. He said that it is more interesting to find a shortest network rather than a point. Gauss also presented several possible connections of the shortest network for four given points.

Unfortunately, Gauss' letter was discovered only in 1986. From 1941 to 1986, many publications have followed R. Courant and H. Robbins who in their popular book [2] called Gauss' problem as the Steiner tree problem. The Steiner tree has become a popular and important name. If you search 'Steiner tree' with 'yahoo.com' on the internet, then you will receive a list of 4675 webpages on Steiner trees. We have no way to change back the name from Steiner trees to Gauss trees. It may be worth mentioning that J. Steiner, a geometrician in 19th century whose name is used for the shortest networks, has not been found so far to have any significant contribution to Steiner trees.

G.B. Dantzig, who first proposed the simplex method to solve linear programming in 1947, stated in [4]: 'What seems to characterize the pre- 1947 era was lack of any interests in trying to optimize'. Due to the lack of interests in optimization, many important works appeared before 1947 were ignored. This happened not only for Steiner trees, but also to other areas of optimization, including some important contributions in linear and nonlinear programming.

The discovery of linear programming started a new age of optimization. However, in [4], Dantzig made the following comment: 'Linear programming was unknown prior to 1947'. This is not quite correct; there were some late exceptions. J.B.J. Fourier (of Fourier series fame) in 1823 and the well-known Belgian mathematician Ch. de la Vallée Poussin in 1911 each wrote a paper about it. Their work had as much influence on post- 1947 developments as would finding in an Egyptian tomb an electronic computer built in 3000 BC. L.V. Kantorovich's remarkable 1939 monograph on the subject was also neglected for ideological reasons in the USSR. It was resurrected two decades later after the major developments had already taken place in the West. An excellent paper by F.L. Hitchcock in 1941 on the transportation problem was also overlooked until after others in the late 1940s and early 1950s have independently rediscovered its properties.

He also recalled how he made his discovery: 'My own contribution grew out of my World War II experience in the Pentagon. During the war period (1941–1945), I had become an expert on programming-planning methods using desk calculators. In 1946 I was mathematical advisor to the US Air Force Comptroller in the Pentagon. I had just received my PhD (for research I had done mostly before the war) and was looking for an academic position that would pay better than a low offer I had received from Berkeley. In order to entice me to not take another job, my Pentagon colleagues, D. Hitchcock and M. Wood, challenged me to see what I could do to mechanize the planning process. I was asked to find a way to more rapidly compute a time-staged development, training and logistical supply program. In those days *mechanizing* planning meant using analog devices or punch-card equipment. There were no electronic computers'.

This challenge problem made Dantzig discover his great work in linear programming without electronic computer. But, we have to point out that it is due to the rapid development of computer technology that applications of linear programming can be made so wide and so great, and areas of optimization can have so fast growing.

In 1951, A.W. Tucker and his student H.W. Kuhn published the Kuhn–Tucker conditions. This is considered as an initial point of nonlinear programming. However, A. Takayama has an interesting comment on these condition: 'Linear programming aroused interest in constraints in the form of inequalities and in the theory of linear inequalities and convex sets. The Kuhn–Tucker study appeared in the middle of this interest with a full recognition of such developments. However, the theory of nonlinear programming when constraints are all in the form of equalities has been known for a long time – in fact, since Euler and Lagrange. The inequality constraints were treated in a fairly satisfactory manner already in 1939 by Karush. Karush's work is apparently under the influence of a similar work in the calculus of variations by Valentine. Unfortunately, Karush's work has been largely ignored'. Yet, this is another work that appeared before 1947 and it was ignored. In the 1960s, G. Zoutendijk, J.B. Rosen, P. Wolfe, M.J.D. Powell, and others published a number of algorithms for solving nonlinear optimization problems. These algorithms form the basis of contemporary nonlinear programming.

In 1954, L.R. Ford and D.R. Fulkerson initiated the study on network flows. This is considered as a starting point on *combinatorial optimization* although Fermat is the first one who studied a major combinatorial optimization problem. In fact, it was because of the influence of the results of Ford and Fulkerson, that interests on combinatorial optimization were growing, and so many problems, including Steiner trees, were proposed or re-discovered in history. In 1958, R.E. Gomory published the cutting plane method. This is considered as an initiation of *integer programming*, an important direction of combinatorial optimization.

In 1955, Dantzig published his paper [3] and E.M.L. Beale proposed an algorithm to solve similar problems. They started the study on *stochastic programming*. R.J-B. Wets in the 1960s, and J.R. Birge and A. Prékopa in the 1980s made important contributions in this branch of optimization.

Now, optimization has merged into almost every corner of economics. New branches of optimization appeared in almost every decade, *global optimization*, *nondifferential optimization*, *geometric programming*, *large scale optimization*, etc. No one in his/her whole life is able to study all branches in optimization. Each researcher can only be an expert in a few branches of optimization.

Of course, the rapid development of optimization is accomplished with recognition of its achievements. One important fact is that several researchers in optimization have received the Nobel Prize in economics, including Kantorovich and T.C. Koopmans. They received the Nobel Prize on economics in 1975 for their contributions to the theory of optimum allocation of resources. H.M. Markowitz received the Nobel Prize on economics in 1990 for his contribution on the quadratic programming model of financial analysis.

Today, optimization has become a very large and important interdisciplinary area between mathematics, computer science, industrial engineering, and management science. The 'International Symposium on Mathematical Programming' is one of major conferences on optimization. From the growing number of papers presented in this conference we may see the projection of growing optimization area:

1949) Chicago, USA, 34 papers;
1951) Washington DC, USA, 19 papers;
1955) Washington DC, USA, 33 papers;
1959) Santa Monica, USA, 57 papers;
1962) Chicago, USA, 43 papers;
1964) London, UK, 83 papers;
1967) Princeton, USA, 91 papers;
1970) The Hague, The Netherlands, 137 papers;
1973) Stanford, USA, about 250 papers;
1976) Budapest, Hungary, 327 papers;
1979) Montreal, Canada, 458 papers;
1982) Bonn, FRG, 554 papers;
1985) Cambridge, USA, 589 papers;
1988) Tokyo, Japan, 624 papers.
(This data is quoted from [1].)

With the current fast growth of computer technology optimization it is expected to continue its great speed of developments. These developments may contain include a deep understanding of the successful heuristics for combinatorial optimization problems with nonlinear programming approaches. It may also include digital simulations to some natural optimization process. As many mysteries and open problems still exist in optimization, it will still be an area receiving a great attention.

See also

- ► Carathéodory, Constantine
- ► Carathéodory Theorem
- ► Inequality-constrained Nonlinear Optimization
- ► Kantorovich, Leonid Vitalyevich
- ► Leibniz, Gottfried Wilhelm
- ► Linear Programming
- ► Operations Research
- ► Von Neumann, John

References

1. Balinski ML (1991) Mathematical programming: Journal, society, recollections. In: Lenstra JK, Rinnooy Kan AHG, Schrijver A (eds) History of Mathematical Programming. North-Holland, Amsterdam, pp 5–18
2. Courant R, Robbins H (1941) What is mathematics? Oxford Univ. Press, Oxford
3. Dantzig GB (1955) Linear programming under uncertainty. Managem Sci 1:197–206
4. Dantzig GB (1991) Linear programming: The story about how it began. In: Lenstra JK, Rinnooy Kan AHG, Schrijver A (eds) History of Mathematical Programming. North-Holland, Amsterdam, pp 19–31
5. de Fermat P (1934) Abhandlungen über Maxima und Minima. In: Oswalds Klassiker der exakten Wissenschaft, vol 238. H. Miller, reprint from original

Homogeneous Selfdual Methods for Linear Programming

ERLING D. ANDERSEN
Odense University, Odense M, Denmark

MSC2000: 90C05

Article Outline

Keywords

Optimization; Linear programming; Interior point methods; Homogeneous; Selfdual

The linear program

$$\begin{cases} \min & c^\top x \\ \text{s.t.} & Ax = b, \\ & x \geq 0 \end{cases} \tag{1}$$

may have an optimal solution, be primal infeasible or be dual infeasible for a particular set of data $c \in \mathbf{R}^n$, $b \in \mathbf{R}^m$, and $A \in \mathbf{R}^{m \times n}$. In fact the problem can be both primal and dual infeasible for some data where (1) is denoted dual infeasible if the dual problem

$$\begin{cases} \max & b^\top y \\ \text{s.t.} & A^\top y + s = c, \\ & s \geq 0 \end{cases} \tag{2}$$

corresponding to (1) is infeasible. The vector s is the so-called *dual slacks*.

However, most methods for solving (1) assume that the problem has an optimal solution. This is in particular true for interior point methods. To overcome this problem it has been suggested to solve the *homogeneous and selfdual model*

$$\begin{cases} \min & 0 \\ \text{s.t.} & Ax - b\tau = 0, \\ & -A^\top y + c\tau \geq 0, \\ & b^\top y - c^\top x \geq 0, \\ & x \geq 0, \quad \tau \geq 0, \end{cases} \tag{3}$$

instead of (1). Clearly, (3) is a homogeneous LP and is selfdual which essentially follows from the constraints form a skew-symmetric system. The interpretation of (3) is τ is a homogenizing variable and the constraints represent primal feasibility, dual feasibility, and reversed weak duality.

The homogeneous model (3) was first studied by A.J. Goldman and A.W. Tucker [2] in 1956 and they proved that (3) always has a nontrivial solution (x^*, y^*, τ^*) satisfying

$$\begin{cases} x_j^* s_j^* = 0, & \forall j \\ x_j^* + s_j^* > 0, & \forall j, \\ \tau^* \kappa^* = 0, \\ \tau^* + \kappa^* > 0, \end{cases} \tag{4}$$

where $s^* := c\,\tau^* - A^\top y^* \geq 0$ and $\kappa^* := b^\top y^* - c^\top x^* \geq 0$. A solution to (3) satisfying the condition (4) is said to be a *strictly complementary solution*. Moreover, Goldman and Tucker showed that if $(x^*, \tau^*, y^*, s^*, \kappa^*)$ is any strictly complementary solution, then exactly one of the two following situations occurs:

- $\tau^* > 0$ if and only if (1) has an optimal solution. In this case$(x^*, y^*, s^*)/\tau^*$ is an optimal primal-dual solution to (1).
- $\kappa^* > 0$ if and only if (1) is primal or dual infeasible. In the case $b^\top y^* > 0$ ($c^\top x^* < 0$) then (1) is primal (dual) infeasible.

The conclusion is that a strictly complementary solution to (3) provides all the information required, because in the case $\tau^* > 0$ then an optimal primal-dual solution to (1) is trivially given by $(x, y, s) = (x^*, y^*, s^*)/\tau^*$. Otherwise, the problem is primal or dual infeasible. Therefore, the main algorithmic idea is to compute a strictly complementary solution to (3) instead of solving (1) directly.

Y. Ye, M.J. Todd, and S. Mizuno [6] suggested to solve (3) by solving the problem

$$\begin{cases} \min & n^0 z \\ \text{s.t.} & Ax - b\tau - \bar{b}z = 0, \\ & -A^\top y + c\tau + \bar{c}z \geq 0, \\ & b^\top y - c^\top x + \bar{d}z \geq 0, \\ & \bar{b}^\top y - \bar{c}^\top x - \bar{d}\tau = -n^0, \\ & x \geq 0, \quad \tau \geq 0, \end{cases} \tag{5}$$

where

$$\bar{b} := Ax^0 - b\tau^0,$$
$$\bar{c} := -c\tau^0 + A^\top y^0 + s^0,$$
$$\bar{d} := c^\top x^0 - b^\top y^0 + \kappa^0,$$
$$n^0 := (x^0)^\top s^0 + \tau^0 \kappa^0,$$

and

$$(x^0, \tau^0, y^0, s^0, \kappa^0) = (e, 1, 0, 1)$$

(e is an n vector of all ones). It can be proved that the problem (5) always has an optimal solution. Moreover, the optimal value is identical to zero and it is easy to verify that if (x, τ, y, z) is an optimal strictly complementary solution to (5), then (x, τ, y) is a strictly complementary solution to (3). Hence, the problem (5) can solved using any method that generates an optimal strictly complementary solution because the problem always has a solution. Note by construction then $(x, \tau, y, z) = (x^0, \tau^0, y^0, 1)$ is an interior feasible solution to (5). This implies that the problem (1) can be solved by most feasible-interior point algorithms.

X. Xu, P.-F. Hung, and Ye [4] suggest an alternative solution method which is also an interior point algorithm, but specially adapted to the problem (3).The so-called *homogeneous algorithm* can be stated as follows:

1) Choose $(x^0, \tau^0, y^0, s^0, \kappa^0)$ such that $(x^0, \tau^0, s^0, \kappa^0) > 0$. Choose $\varepsilon_f, \varepsilon_g > 0$ and $\gamma \in (0, 1)$ and let $\eta := 1 - \gamma$.
2) $k := 0$.
3) Compute:

$$
\begin{aligned}
r_p^k &:= b\tau^k - Ax^k,\\
r_d^k &:= c\tau^k - A^\top y^k - s^k,\\
r_g^k &:= \kappa^k + c^\top x^k - b^\top y^k,\\
\mu^k &:= \frac{(x^k)^\top s^k + \tau^k\kappa^k}{n+1}.
\end{aligned}
$$

4) If $\| (r_p^k; r_d^k; r_g^k) \| \le \varepsilon_f$ and $\mu^k \le \varepsilon_g$, then terminate.
5) Solve the linear equations

$$
\begin{aligned}
Ad_x - bd_\tau &= \eta r_p^k,\\
A^\top d_y + d_s - cd_\tau &= \eta r_d^k,\\
-c^\top d_x + b^\top d_y - d_\kappa &= \eta r_g^k,\\
S^k d_x + X^k d_s &= -X^k s^k + \gamma\mu^k e,\\
\kappa^k d_\tau + \tau^k d_\kappa &= -\tau^k\kappa^k + \gamma\mu^k,
\end{aligned}
$$

for $(d_x, d_\tau, d_y, d_s, d_\kappa)$ where $X^k := \mathrm{diag}(x^k)$ and $S^k := \mathrm{diag}(s^k)$.
6) For some $\theta \in (0, 1)$, let α^k be the optimal objective value to

$$
\begin{cases}
\max & \theta\alpha\\
\text{s.t.} & \begin{pmatrix} x^k\\ \tau^k\\ s^k\\ \kappa^k \end{pmatrix} + \alpha \begin{pmatrix} d_x\\ d_\tau\\ d_s\\ d_\kappa \end{pmatrix} \ge 0,\\
& \alpha \le \theta^{-1}.
\end{cases}
$$

7)

$$
\begin{pmatrix} x^{k+1}\\ \tau^{k+1}\\ y^{k+1}\\ s^{k+1}\\ \kappa^{k+1} \end{pmatrix} := \begin{pmatrix} x^k\\ \tau^k\\ y^k\\ s^k\\ \kappa^k \end{pmatrix} + \alpha^k \begin{pmatrix} d_x\\ d_\tau\\ d_y\\ d_s\\ d_\kappa \end{pmatrix}
$$

8) $k = k + 1$.
9) goto 3)

The following facts can be proved about the algorithm

$$
\begin{cases}
r_p^{k+1} = (1 - (1-\gamma)\alpha^k) r_p^k,\\
r_d^{k+1} = (1 - (1-\gamma)\alpha^k) r_d^k,\\
r_g^{k+1} = (1 - (1-\gamma)\alpha^k) r_g^k,
\end{cases}
$$

and

$$
\begin{aligned}
&((x^{k+1})^\top s^{k+1} + \tau^{k+1}\kappa^{k+1})\\
&= (1 - (1-\gamma)\alpha^k)((x^k)^\top s^k + \tau^k\kappa^k),
\end{aligned}
$$

which shows that the primal residuals (r_p), the dual residuals (r_d), the gap residual (r_g), and the complementary gap $(x^\top s + \tau\kappa)$ all are reduced strictly if $\alpha^k > 0$ and at the same rate. This shows that $(x^k, \tau^k, y^k, s^k, \kappa^k)$ generated by the algorithm converges towards an optimal solution to (3) (and the termination criteria in step 4) is ultimately reached). In principle the initial point and the stepsize α^k should be chosen such that

$$
\min_j (x_j^k s_j^k, \tau^k\kappa^k) \ge \beta\mu^k, \quad \text{for } k = 0, 1, \ldots,
$$

is satisfied for some $\beta \in (0, 1)$ because this guarantees $(x^k, \tau^k, y^k, s^k, \kappa^k)$ converges towards a strictly complementary solution. Finally, it is possible to prove that the algorithm has the complexity $O(n^{3.5}L)$ given an appropriate choice of the starting point and the algorithmic parameters.

Further details about the homogeneous algorithm can be seen in [3,5]. Issues related to implementing the homogeneous algorithm are discussed in [1,4].

See also

- ▶ Entropy Optimization: Interior Point Methods
- ▶ Interior Point Methods for Semidefinite Programming
- ▶ Linear Programming: Interior Point Methods

- ▶ Linear Programming: Karmarkar Projective Algorithm
- ▶ Potential Reduction Methods for Linear Programming
- ▶ Sequential Quadratic Programming: Interior Point Methods for Distributed Optimal Control Problems
- ▶ Successive Quadratic Programming: Solution by Active Sets and Interior Point Methods

References

1. Andersen ED, Andersen KD (2000) The MOSEK interior point optimizer for linear programming: An implementation of the homogeneous algorithm. In: Frenk H, Roos K, Terlaky T, Zhang S (eds) High Performance Optimization. Kluwer, Dordrecht, pp 197–232
2. Goldman AJ, Tucker AW (1956) Theory of linear programming. In: Kuhn HW, Tucker AW (eds) Linear Inequalities and related Systems. Princeton Univ. Press, Princeton, pp 53–97
3. Roos C, Terlaky T, Vial J-P (1997) Theory and algorithms for linear optimization: An interior point approach. Wiley, New York
4. Xu X, Hung P-F, Ye Y (1996) A simplified homogeneous and self-dual linear programming algorithm and its implementation. Ann Oper Res 62:151–171
5. Ye Y (1997) Interior point algorithms: theory and analysis. Wiley, New York
6. Ye Y, Todd MJ, Mizuno S (1994) An $O(\sqrt{n}L)$-iteration homogeneous and self-dual linear programming algorithm. Math Oper Res 19:53–67

Hyperplane Arrangements

PETER ORLIK
Department Math., University Wisconsin, Madison, USA

MSC2000: 52C35, 05B35, 57N65, 20F36, 20F55

Article Outline

Keywords

Hyperplane arrangement; Geometric semilattice; Orlik–Solomon algebra; Divisor; Singularity; Complement; Homotopy type; Poincaré polynomial; Ball quotient; Logarithmic form; Hypergeometric integral

Let V be an ℓ-dimensional affine space over the field $\mathbf{K}$. An *arrangement of hyperplanes*, $\mathcal{A}$, is a finite collection of codimension one affine subspaces in V, [5].

Some Examples

1) A subset of the coordinate hyperplanes is called a *Boolean arrangement*.
2) An arrangement is in *general position* if at each point it is locally Boolean.
3) The *braid arrangement* consists of the hyperplanes $\{x_i = x_j : 1 \le i < j \le \ell\}$. It is the set of reflecting hyperplanes of the symmetric group on ℓ letters.
4) The reflecting hyperplanes of a finite reflection group is a *reflection arrangement*.

Combinatorics

An edge X of $\mathcal{A}$ is a nonempty intersection of elements of $\mathcal{A}$. Let $L(\mathcal{A})$ be the set of edges partially ordered by reverse inclusion. Then L is a *geometric semilattice* with minimal element V, rank given by codimension, and maximal elements of the same rank, $r(\mathcal{A})$. The *Moebius function* on L is defined by $\mu(V) = 1$ and for $X > V$, $\sum_{V \le Y \le X} \mu(Y) = 0$. The *characteristic polynomial* of $\mathcal{A}$ is $\chi(\mathcal{A}, t) = \sum_{X \in L} \mu(X) t^{\dim X}$. The β-invariant of $\mathcal{A}$ is $\beta(\mathcal{A}) = (-1)^{r(\mathcal{A})} \chi(\mathcal{A}, 1)$. For a generic arrangement of n hyperplanes $\chi(\mathcal{A}, t) = \sum_{k=0}^{r(\mathcal{A})} (-1)^k \binom{n}{k} t^{\ell - k}$. For the braid arrangement $\chi(\mathcal{A}, t) = t(t-1)(t-2)\cdots(t-(\ell-1))$. Similar factorizations hold for all reflection arrangements involving the (co)exponents of the reflection group. Given a p-tuple of hyperplanes, $S = (H_1, \ldots, H_p)$, let $\cap S = H_1 \cap \cdots \cap H_p$ and note that $\cap S$ may be empty. We say that S is *dependent* if $\cap S \neq \emptyset$ and $\operatorname{codim}(\cap S) < |S|$. Let $E(\mathcal{A})$ be the exterior algebra on symbols (H) for $H \in \mathcal{A}$ where product is juxtaposition. Define $\partial : E \to E$ by $\partial 1 = 0$, $\partial(H) = 1$ and for $p \ge 2$, $\partial(H_1 \cdots H_p) = \sum_{k=1}^{p} (-1)^{k-1} (H_1 \cdots \widehat{H_k} \cdots H_p)$. Let $I(\mathcal{A})$ be the ideal of $E(\mathcal{A})$ generated by $\{S : \cap S = \emptyset\} \cup \{\partial S : S \text{ is dependent}\}$. The *Orlik–Solomon algebra* of $\mathcal{A}$ is

$A(\mathcal{A}) = E(\mathcal{A})/I(\mathcal{A})$. See also connections with *matroid theory* [3].

Divisor

The *divisor* of $\mathcal{A}$ is the union of the hyperplanes, $N(\mathcal{A})$. If $\mathbf{K} = \mathbf{R}$ or $\mathbf{K} = \mathbf{C}$, then N has the *homotopy type* of a wedge of $\beta(\mathcal{A})$ spheres of dimension $r(\mathcal{A}) - 1$, [4]. The *singularities* of N are not isolated. The divisor of a general position arrangement has normal crossings, but this is not true for arbitrary $\mathcal{A}$. Blowing up N along all edges where it is not locally a product of arrangements yields a normal crossing divisor.

Complement

The *complement* of $\mathcal{A}$ is $M(\mathcal{A}) = V - N(\mathcal{A})$.

1) If $\mathbf{K} = \mathbf{F}_q$, then M is a finite set of cardinality $|M| = \chi(\mathcal{A}, q)$.
2) If $\mathbf{K} = \mathbf{R}$, then M is a disjoint union of open convex sets (*chambers*) of cardinality $(-1)^\ell \chi(\mathcal{A}, -1)$. If $r(\mathcal{A}) = \ell$, M contains $\beta(\mathcal{A})$ chambers with compact closure, [7].
3) If $\mathbf{K} = \mathbf{C}$, then M is an open complex (Stein) *manifold* of the homotopy type of a finite CW complex.

Its *cohomology* is torsion-free and its *Poincaré polynomial* is $\mathrm{Poin}(M, t) = (-t)^\ell \chi(\mathcal{A}, -t^{-1})$. The product structure is determined by the isomorphism of graded algebras $H^*(M) \simeq A(\mathcal{A})$. The *fundamental group* of M has an effective presentation but the *higher homotopy groups* of M are not known in general. The complement of a Boolean arrangement is a complex torus. In a general position arrangement of $n > \ell$ hyperplanes M has nontrivial higher homotopy groups. For the braid arrangement, M is called the pure braid space and its higher homotopy groups are trivial. The symmetric group acts freely on M with orbit space the braid space whose fundamental group is the *braid group*. The quotient of the divisor by the symmetric group is called the *discriminant*, which has connections with singularity theory.

Ball Quotients

Examples of algebraic surfaces whose *universal cover* is the complex ball were constructed as 'Kummer' covers of the projective plane branched along certain arrangements of projective lines, [2].

Logarithmic Forms

For $H \in \mathcal{A}$ choose a linear polynomial α_H with $H = \ker \alpha_H$ and let $Q(\mathcal{A}) = \prod_{H \in \mathcal{A}} \alpha_H$. Let $\Omega^p[V]$ denote all global regular (i. e., polynomial) p-forms on V. Let $\Omega^p(V)$ denote the space of all global rational p-forms on V. The space $\Omega^p(\mathcal{A})$ of logarithmic p-forms with poles along $\mathcal{A}$ is

$$\Omega^p(\mathcal{A}) = \{\omega \in \Omega^p(V) \colon\ Q\omega \in \Omega^p[V], \\ Q(d\omega) \in \Omega^{p+1}[V]\}\,.$$

The arrangement is free if $\Omega^1(\mathcal{A})$ is a free module over the polynomial ring. A *free arrangement* $\mathcal{A}$ has integer exponents $\{b_1, \ldots, b^\ell\}$ so that $\chi(\mathcal{A}, t) = \prod_{k=1}^{\ell}(t - b_k)$. Reflection arrangements are free. This explains the factorization of their characteristic polynomials.

Hypergeometric Integrals

Certain rank one *local system cohomology* groups of M may be identified with spaces of *hypergeometric integrals*, [1]. If the local system is suitably generic, these cohomology groups may be computed using the algebra $A(\mathcal{A})$. Only the top cohomology group is nonzero and it has dimension $\beta(\mathcal{A})$. See [6] for connections with the representation theory of *Lie algebras* and *quantum groups*, and with the *Knizhnik–Zamolodchikov differential equations* of physics.

See also

► Hyperplane Arrangements in Optimization

References

1. Aomoto K, Kita M (1994) Hypergemetric functions. Springer, Berlin
2. Barthel G, Hirzebruch F, Höfer T (1987) Geradenkonfigurationen und Algebraische Flächen. Vieweg, Braunschweig/Wiesbaden
3. Björner A, Las Vergnas M, Sturmfels B, White N, Ziegler GM (1993) Oriented matroids. Cambridge Univ. Press, Cambridge
4. Goresky M, MacPherson R (1988) Stratified Morse theory. Springer, Berlin
5. Orlik P, Terao H (1992) Arrangements of hyperplanes. Springer, Berlin
6. Varchenko A (1995) Multidimensional hypergeometric functions and representation theory of Lie algebras and quantum groups. World Sci., Singapore

7. Zaslavsky T (1975) Facing up to arrangements: Face-count formulas for partitions of space by hyperplanes. Memoirs Amer Math Soc 154

Hyperplane Arrangements in Optimization

PANOS M. PARDALOS
Center for Applied Optim. Department Industrial and Systems Engineering, University Florida, Gainesville, USA

MSC2000: 05B35, 20F36, 20F55, 52C35, 57N65

Article Outline

Keywords

Hyperplane arrangement; Polynomial time algorithm

A finite set S of hyperplanes in $\mathbf{R}^d$ defines a dissection of $\mathbf{R}^d$ into connected sets of various dimensions. We call this dissection the *arrangement* $A(S)$ of S.

Given a vector $\eta = (\eta_1, \ldots, \eta_d) \in \mathbf{R}^d - \{0\}$ and a number $\eta_0 \in \mathbf{R}$, we may define a hyperplane H and associated halfspaces H^-, H^+ by

$$H = \left\{x \in \mathbf{R}^d : \ \eta \cdot x = \eta_0\right\},$$
$$H^- = \left\{x \in \mathbf{R}^d : \ \eta \cdot x < \eta_0\right\},$$
$$H^+ = \left\{x \in \mathbf{R}^d : \ \eta \cdot x > \eta_0\right\}.$$

Clearly, H, H^-, H^+ are disjoint and $H \cup H^- \cup H^+ = \mathbf{R}^d$.

We may now specify the location of a point relative to the set of hyperplanes $S = \{H_1, \ldots, H_n\}$. For a point p and $1 \le j \le n$, define

$$s_j(p) = \begin{cases} -1 & \text{if } p \in H_j^-, \\ 0 & \text{if } p \in H_j, \\ +1 & \text{if } p \in H_j^+. \end{cases}$$

The vector $s(p) = (s_1(p), \ldots, s_n(p))$ is called the *position vector* of p.

Clearly there are at most 3^n possible position vectors, however, in general most of these will not occur. We say that points p and q *lie on the same face* if $s(p) = s(q)$. The nonempty set of points with position vector r is called the *face* $f(r)$:

$$f(r) = \left\{p \in \mathbf{R}^d : \ s(p) = r\right\}$$

The nonempty sets of this form are called the *faces of the arrangement* $A(S)$. The position vector of a face $f(r) = g$ is defined to be r,

$$s(f(r)) = r.$$

A face f is called a k-face if its dimension is k. Special names are used to denote k-faces for special values of k: a 0-face is called a *vertex*, a 1-face is called an *edge*, a $(d-1)$-face is called a *facet*, and a d-face is called a *cell*. A face is said to be a *subface* of another face g if the dimension of f is one less than the dimension of g and f is contained in the boundary of g; it follows that $s_i(f) = 0$ unless $s_i(f) = s_i(g)$ for $1 \le i \le n$. If f is a subface of g, then we also say that f and g are *incident* (upon each other) or that they define an *incidence*.

An arrangement $A(S)$ of $n \ge d$ hyperplanes is called *simple* if any d hyperplanes of S have a unique point in common and if any $d+1$ hyperplanes have no point in common. If $n < d$, we say that $A(S)$ is simple if the common intersection of the n hyperplanes is a $(d-n)$-flat. For more details see [3,4] and [5].

As an application of hyperplane arrangements in algorithm design for optimization problems, see [1].In it the problem of minimizing the Euclidean distance function on $\mathbf{R}^n$ subject to m equality constraints and upper and lower bounds (box constraints) is considered. A parametric characterization in $\mathbf{R}^m$ of the family of solutions to this problem is provided, thereby showing equivalence with a problem of search in an arrangement of hyperplanes in $\mathbf{R}^m$. This characterization and the technique for constructing arrangements due to H. Edelsbrunner, J. O'Rourke and R. Seidel are used to develop an exact algorithm for the problem. The algo-

rithm is strongly polynomial running in time $\Theta(n^m)$ for each fixed m.

See also

► Hyperplane Arrangements

References

1. Berman P, Kovoor N, Pardalos PM (1993) Algorithms for the least distance problem. Complexity in Numerical Optimization. World Sci., Singapore, pp 33–56
2. Chazelle B, Guibas J, Lee DT (1985) The power of geometric duality. Proc. 15th Annual ACM Symp. Theory of Computing. ACM, New York, pp 217–225
3. Edelsbrunner H (1987) Algorithms in combinatorial geometry. Springer, Berlin
4. Edelsbrunner H, O'Rourke J, Seidel R (1986) Constructing arrangements of lines and hyperplanes with applications. SIAM J Comput 15:341–363
5. Orlik P, Terao H (1992) Arrangements of hyperplanes. Springer, Berlin
6. Pardalos PM, Kovoor N (1990) An algorithm for a singly constrained class of quadratic programs subject to upper and lower bounds. Math Program 46:321–328