JOURNAL OF OPTIMIZATION THEORY AND APPLICATIONS: Vol. 37, No. 2, JUNE 1982

Minimizing a Differentiable Function over a Differential Manifold[1]

D. GABAY[2]

Communicated by D. G. Luenberger

Abstract. To generalize the descent methods of unconstrained optimization to the constrained case, we define intrinsically the gradient field of the objective function on the constraint manifold and analyze descent methods along geodesics, including the gradient projection and reduced gradient methods for special choices of coordinate systems. In particular, we generalize the quasi-Newton methods and establish their superlinear convergence; we show that they only require the updating of a reduced size matrix. In practice, the geodesic search is approximated by a tangent step followed by a constraints restoration or by a simple arc search again followed by a restoration step.

Key Words. Nonlinearly constrained optimization, differential geometry, Riemannian manifolds, iterative methods, convergence theorems, rate of convergence.

1. Introduction

This paper shares with a previous article by the author and Luenberger (Ref. 1) the purpose of extending the well-known gradient-related methods for the unconstrained minimization of a real-valued function on $\mathbb{R}^n$ to the nonlinearly constrained problem

$$\min\{f(x) \mid x \in \mathbb{R}^n : c_i(x) = 0,\ i = 1, 2, \ldots, m\}, \tag{1}$$

where $m \leq n$, f and c_i are real-valued functions on $\mathbb{R}^n$ and assumed to be $\mathscr{C}^\sigma$ differentiable ($\sigma \geq 2$, unless otherwise specified). Denoting by c the

[1] A first draft of this paper was first presented at the 10th International Symposium on Mathematical Programming, Montreal, Canada, 1979. The present version has been prepared while the author was visiting the Department of Engineering-Economic Systems at Stanford University, partially supported by AFOSR Grant No. 77-3141.

[2] Research Associate, Laboratoire d'Analyse Numérique, Université P. et M. Curie, Paris, France; and Scientific Advisor, INRIA, Domaine de Voluceau, Le Chesnay, France.

0022-3239/82/0600-0177$03.00/0

map from $\mathbb{R}^n$ into $\mathbb{R}^m$ of component c_i, we also assume that the following *regularity assumption* holds: 0 is a regular value of the map c; i.e., the Jacobian $c'(x) \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$ of c at x is of full rank m for all $x \in C = c^{-1}(0)$.

In Ref. 1, this program was fulfilled in the following way. The regularity assumption implies that there exists an $m \times m$ submatrix of the Jacobian matrix A_x of c at $x \in C$ which is nonsingular; say,

$$B = (\partial c_i / \partial x_j),\ i, j \in I = \{1, \ldots, m\},$$

and let

$$J = N - I.$$

Identify $\mathbb{R}^n$ with $\mathbb{R}^m \times \mathbb{R}^{n-m}$, and set

$$x = (x_I, x_J),\ A_x = [B, D].$$

According to the implicit function theorem, there exist a neighborhood $U_x = W \times V$ of $x = (x_I, x_J)$ in $\mathbb{R}^m \times \mathbb{R}^{n-m}$ and a $\mathcal{C}^\sigma$ map $\psi_x : V \to W$ such that

$$y_I = \psi_x(y_J), \qquad \text{iff } c(y_I, y_J) = 0.$$

The restriction of f to $U_x \cap C$ can thus be represented by a $\mathcal{C}^\sigma$ real-valued function $\varphi_x : V \subset \mathbb{R}^{n-m} \to \mathbb{R}$ defined by

$$\varphi_x(z) = f(\psi_x(z), z). \tag{2}$$

Suppose now that $x^* \in C$ is a local minimizer of f on C in the neighborhood $U_x \cap C$ of x. Then,

$$x^* = (x_I^*, x_J^*), \qquad \text{with } x_I^* = \psi_x(x_J^*),$$

and x_J^* is a local minimizer of the (unconstrained) minimization problem on the open set V

$$\min_{z \in V} \varphi_x(z). \tag{3}$$

Applying the gradient-related methods to the unconstrained reduced problem (3), we construct a sequence of approximations $\{z^k\}$ converging to x_J^* defined iteratively by

$$z^{k+1} = z^k + t_k p^k, \tag{4}$$

where p^k is a descent direction based on g^k the gradient of φ_x at z^k, and t_k is a stepsize selected by a line search of the function

$$j(t) = \varphi_x(z^k + t p^k), \qquad \text{for } t > 0.$$

In the original space $\mathbb{R}^n$, iteration (4) corresponds to the scheme

$$x_J^{k+1} = x_J^k + t_k p^k, \tag{5a}$$

$$x_I^{k+1} = \psi(x_J^{k+1}). \tag{5b}$$

The gradient of φ_x can be expressed in terms of the original data of the problem

$$g^k = \nabla_J f(x^k) - D^T (B^{-1})^T \nabla_I f(x^k) \tag{6}$$

and is called the *reduced gradient* for the partition $N = I \oplus J$; the stepsize t_k must be selected by a search along the arc of curve starting from x^k given by

$$x(t) = (\psi(x_J^k + tp^k), x_J^k + tp^k) \tag{7}$$

to produce a sufficient decrease of the objective function. With this framework, Gabay and Luenberger proposed idealized methods based on the reduced gradient extending the steepest descent method, Newton's method, and a quasi-Newton method and analyzed their convergence properties. Notice that, if $x^* \notin U_{x^k} \cap C$, it is still possible to find a new approximation x^{k+1} according to (5) by restricting the search along the curve (7) to an interval $[0, \bar{t}]$ such that $x^{k+1} \in U_{x^k} \cap C$; at x^{k+1}, a new partition $N = I \oplus J$ can be constructed, and an iterate x^K is eventually reached such that $x^* \in U_{x^K} \cap C$.

Some problems possess an underlying structure which leads to a natural partition of the variables; for instance, in optimal control problems (considered as constrained by the dynamics equation), an obvious choice consists in distinguishing between control and state variables: ψ then maps the sequence of controls into the resulting trajectory in the state space and (5b) can be performed by an integration of the dynamics equation. However, such structural guidelines are not always available, and arbitrary partitions may lead to poor computational performances.

In this paper, we extend the approach of Ref. 1 in order to define a family of methods which exploit the intrinsic reduction of dimensionality of the problem, but avoid any unnatural asymmetric treatment of the variables. In addition, the new proposed framework allows us to embody in the methods insights about the problem structure which enhance their numerical performances.

This extension is achieved by exploiting the differential manifold structure of the constrained set C, whereas the analysis of Ref. 1 was confined to the particular local coordinate system induced by the partition. Using this geometric framework, we can define directly gradient-related methods

for the minimization of the real-valued function f over the differential manifold C, independently of the choice of coordinate systems. Similarly to the unconstrained case, where we (implicitly) use the Euclidean structure of $\mathbb{R}^n$, we must endow C with a *Riemannian structure*. We can then define the gradient (field) of f on C and descent methods consisting in generating from an approximate solution x^k a new iterate x^{k+1} on the *geodesic curve* starting from x^k and tangent to a direction related to the gradient of f on C at x^k. This approach was introduced by Luenberger (Ref. 2) to study the convergence properties of Rosen's gradient projection method (Ref. 3); the Riemannian structure on C used (implicitly) in Ref. 2 is the one induced by the Euclidean structure of $\mathbb{R}^n$. See also Lichnewsky (Ref. 4), where a gradient and a conjugate-gradient method along geodesics is analyzed. We find it important to develop our analysis for a general Riemannian structure, not only because it seems the appropriate theoretical framework, but because it will allow us to choose the Riemannian metric according to (our knowledge of) the nature of specific problems and will result in improved computational performances.

In Section 2, we precisely show that, under the regularity assumption, $C = c^{-1}(0)$ is a differential manifold and introduce a nonlinear change of coordinates in $\mathbb{R}^n$ which is useful for the practical implementation of our methods. We also give an estimate of the diameter of the neighborhood $U_x \cap C$ on which a coordinate system around x can be defined. Section 3 is of a tutorial nature and presents important results on the geometry of a Riemannian manifold which are used in our analysis. Section 4 constitutes the core of the paper. We first establish optimality conditions in terms of the restriction of f to the manifold C and show their relations with the traditional Lagrange multiplier rules. We then define the gradient field of f on C and show how it depends upon the Riemannian metric; we thus define the reduced gradient in a local coordinate system which yields the reduced gradient used in Ref. 1 and Rosen's projected gradient as special cases. We then define the steepest descent method, the Newton's method, and quasi-Newton methods along geodesics and analyze their convergence properties. In Section 5, these methods are specified in the framework of a coordinate system. The analysis of Section 2 then provides a practical scheme for the implementation of the algorithms which can be interpreted as the sequential tangent-restoration approach introduced by Rosen (Ref. 3) and used by Abadie and Guigou (Ref. 5) and by Miele et al. (Ref. 6). We give in particular two efficient versions of a quasi-Newton method for constrained minimization based on a generalized Broyden–Fletcher–Goldfarb–Shanno update formula. Finally, in Section 6, we indicate how our approach can be used to handle problems with inequality constraints.

2. Constrained Set as a Manifold

In this paragraph, we study the properties of the constrained set $C = c^{-1}(0)$, defined by the map $c: \mathbb{R}^n \to \mathbb{R}^m$, $m \leq n$, assumed to be $\mathscr{C}^\sigma$ differentiable, $\sigma \geq 2$. Recall that 0 is a regular value of c iff $c'(x) \in \mathscr{L}(\mathbb{R}^n, \mathbb{R}^m)$ is of full rank m for all $x \in C$.

Theorem 2.1. Let $c: \mathbb{R}^n \to \mathbb{R}^m$, $m \leq n$, be a $\mathscr{C}^\sigma$ map such that 0 is a regular value of c. Then, the set $C = c^{-1}(0)$ is an $(n-m)$-dimensional submanifold of $\mathbb{R}^n$ of class $\mathscr{C}^{\sigma-1}$.

Proof. To prove this result, often referred to as the regular value theorem, we follow Milnor (Ref. 7) and introduce a nonlinear change of coordinates in $\mathbb{R}^n$ which will be useful for our analysis.

Fix a point $x \in C$; thus, $c(x) = 0$. Since 0 is a regular value of c, the derivative $c'(x) \in \mathscr{L}(\mathbb{R}^n, \mathbb{R}^m)$ must map $\mathbb{R}^n$ onto $\mathbb{R}^m$. The nullspace $\mathscr{N}[c'(x)]$ is therefore an $(n-m)$-dimensional subspace of $\mathbb{R}^n$. Now, choose a linear map $Z_x \in \mathscr{L}(\mathbb{R}^n, \mathbb{R}^{n-m})$ that is nonsingular on this subspace, i.e., such that

$$\mathscr{N}[Z_x] \cap \mathscr{N}[c'(x)] = \{0\}. \tag{8}$$

Define the mapping $s_x: \mathbb{R}^n \to \mathbb{R}^m \times \mathbb{R}^{n-m}$ by

$$s_x(y) = \begin{bmatrix} c(y) \\ Z_x(y-x) \end{bmatrix}, \qquad \text{for all } y \in \mathbb{R}^n. \tag{9}$$

Notice that $s_x(x) = 0$ and that the derivative $s_x'(x) \in \mathscr{L}(\mathbb{R}^n, \mathbb{R}^n)$, given by

$$s_x'(x)(y) = \begin{bmatrix} c'(x)(y) \\ Z_x(y) \end{bmatrix}, \qquad \text{for all } y \in \mathbb{R}^n, \tag{10}$$

is nonsingular because of (8).

The inverse function theorem implies that s_x maps some neighborhood U_x of x in $\mathbb{R}^n$ diffeomorphically onto a neighborhood $W \times V$ of 0 in $\mathbb{R}^m \times \mathbb{R}^{n-m}$. Under the change of coordinates s_x, the set C is transformed *locally* in the $(n-m)$-dimensional subspace $\{0\} \times \mathbb{R}^{n-m}$ of $\mathbb{R}^n$, since s_x maps $U_x \cap C$ diffeomorphically onto $\{0\} \times V$. The map Z_x defines by restriction a *coordinate system* of C around x on the coordinate domain $U_x \cap C$, and the components $(z_1, \ldots, z_{n-m})$ of the image $z = Z_x(y-x)$ of $y \in U_x \cap C$ are called the *local coordinates* of y. The inverse diffeomorphism s_x allows to define a *local (nonlinear) parametrization* $\theta_x: V \to U_x \cap C$ of C around x by

$$y = \theta_x(z) = s_x^{-1}(0, z). \tag{11}$$

Thus, C is an $(n-m)$-dimensional manifold. Provided the map Z_x is chosen as a $\mathscr{C}^{\sigma-1}$ differentiable function of x, the change from local coordinates

around x to local coordinates around x' (such that $U_x \cap U_{x'} \cap C \neq \emptyset$), defined by

$$Z_{x'} \circ \theta_x : Z_x(U_x \cap U_{x'} \cap C) \to Z_{x'}(U_x \cap U_{x'} \cap C),$$

is $\mathscr{C}^{\sigma-1}$ differentiable and the manifold C is of class $\mathscr{C}^{\sigma-1}$. □

Since we are mainly interested in computational procedures, we may wish to specify further the choice of the map Z_x defining the coordinate system of C. Let us denote by A_x the $m \times n$ Jacobian matrix of c at x, while we keep Z_x to denote the $(n-m) \times n$ matrix representing the linear map Z_x. The Jacobian matrix of the map s_x is the $n \times n$ matrix S_x given in partitioned form by

$$S_x = \begin{bmatrix} A_x \\ Z_x \end{bmatrix}. \tag{12}$$

Notice that the regularity assumption implies that

$$\operatorname{rank}(A_x) = m, \qquad \text{for all } x \in C.$$

Recall that a *right inverse* M^- for an $l \times n$ matrix M of full rank $l \leq n$ is an $n \times l$ matrix of full rank l such that

$$MM^- = I.$$

Such an inverse exists, but is not unique. The following result relates the choice of Z_x (and Z_x^-) to a particular choice of a right inverse A_x^- of A_x in order to express the inverse S_x^{-1} in a simple form.

Proposition 2.1. Let A_x^- be a right inverse for A_x. Then, there exists an $(n-m) \times n$ matrix Z_x of full rank $n-m$ and a right inverse Z_x^- satisfying

$$Z_x A_x^- = 0, \qquad A_x Z_x^- = 0, \tag{13}$$

such that the matrix S_x given by (12) is nonsingular and its inverse is given by

$$S_x^{-1} = [A_x^-, Z_x^-]. \tag{14}$$

Proof. A necessary and sufficient condition for the nonsingularity of S_x is

$$\mathscr{N}(Z_x) \cap \mathscr{N}(A_x) = \{0\}. \tag{15}$$

Since A_x is of full rank m, $\mathscr{N}(A_x)$ is a subspace of $\mathbb{R}^n$ of dimension $n-m$. It follows from (15) that $\mathscr{N}(Z_x)$ must be a complement of $\mathscr{N}(A_x)$ in $\mathbb{R}^n$, i.e., an m-dimensional subspace; hence, Z_x must be of full rank $n-m$. The right inverse A_x^- being of rank m, its columns span a subspace $\mathscr{R}(A_x^-)$ of

$\mathbb{R}^n$ of dimension m such that

$$\mathcal{N}(A_x) \cap \mathcal{R}(A_x^-) = \{0\},$$

by definition of a right inverse. Hence,

$$\mathcal{N}(Z_x) = \mathcal{R}(A_x^-),$$

yielding

$$Z_x A_x^- = 0.$$

A similar argument shows that

$$\mathcal{R}(Z_x^-) = \mathcal{N}(A_x), \qquad \text{i.e., } A_x Z_x^- = 0.$$

Formula (14) is established by observing that

$$y = A_x^- a + Z_x^- z$$

satisfies the equation

$$S_x y = \begin{bmatrix} a \\ z \end{bmatrix}$$

for all $a \in \mathbb{R}^m, z \in \mathbb{R}^{n-m}$; and it is the unique solution, since S_x is nonsingular. □

Define the $n \times n$ matrix P_x by

$$P_x = I - A_x^- A_x. \tag{16}$$

It is the matrix of a projection onto $\mathcal{N}(A_x)$, since $A_x P_x = 0$ and the matrices Z_x and Z_x^- defined by Proposition 2.1 satisfy

$$Z_x^- Z_x = P_x. \tag{17}$$

Example 2.1. *Partitioned Right Inverse.* After possibly a permutation of columns, the matrix A_x can be partitioned into

$$A_x = [B, D], \tag{18}$$

such that B is an $m \times m$ nonsingular matrix. It is easy to verify that

$$A_x^- = \begin{bmatrix} B^{-1} \\ 0 \end{bmatrix} \tag{19}$$

is a right inverse for A_x and that

$$Z_x = [0, I_{n-m}], \qquad Z_x^- = \begin{bmatrix} -B^{-1}D \\ I_{n-m} \end{bmatrix} \tag{20}$$

satisfy Proposition 2.1. In this case, the local coordinate system around x

is formed by some $n-m$ coordinates in $\mathbb{R}^{n-m}$; it is the system used by Gabay and Luenberger (Ref. 1).

From a computational viewpoint, this choice requires the inversion of the $m \times m$ matrix B which can be performed by Gaussian elimination with partial pivoting, in the order of $m^3/3$ multiplications; but the partition selected may produce an ill-conditioned matrix B, even though A_x is not ill-conditioned.

Example 2.2. *QR Pseudo-Inverse.* A classical choice for a right inverse of A_x is the Penrose *pseudo-inverse* A_x^+, which in the case of a full-rank matrix A_x can be expressed as

$$A_x^+ = A_x^T (A_x A_x^T)^{-1}. \tag{21}$$

Notice that the matrix P_x defined by (16) is now

$$P_x = I - A_x^T (A_x A_x^T)^{-1} A_x,$$

the *orthogonal projector* onto $\mathcal{N}(A_x)$ in the Euclidean space $\mathbb{R}^n$. Relations (13) are satisfied by choosing Z_x such that $\mathcal{N}(Z_x)$ is the orthogonal complement of $\mathcal{N}(A_x)$ in $\mathbb{R}^n$ and taking for Z_x^- its pseudo-inverse. An equivalent characterization of Z_x is

$$\mathcal{R}(Z_x^T) = \mathcal{N}(A_x), \tag{22}$$

and a particularly convenient choice satisfying (22) consists in taking for columns of Z_x^T an orthonormal basis of $\mathcal{N}(A_x)$; in this case, notice that

$$Z_x^- = Z_x^T,$$

and therefore

$$Z_x^{-T} Z_x^- = I_{n-m}. \tag{23}$$

From a computational viewpoint, notice that the formula (21) involves the inverse of a matrix of condition number $[\mathcal{K}(A)]^2$ and should be avoided in practice. Besides, while it is theoretically possible to generate the columns of Z_x^T by a Gram–Schmidt orthogonalization method, such a procedure is numerically unstable. It is, however, possible to obtain A_x^+ and Z_x in a numerically efficient way, once A_x is factorized as

$$A_x = \Lambda Q, \tag{24}$$

where Q is an $n \times n$ orthogonal matrix,

$$QQ^T = Q^T Q = I,$$

and Λ an $m \times n$ matrix of the form

$$\Lambda = [L, 0],$$

with L being $m \times m$ lower triangular. This *QR decomposition* can be carried out in a numerically stable way using Householder's transformations with approximately $(n - m/3)m^2$ multiplications [see, e.g., Stewart (Ref. 8)].

Partition Q into

$$Q = \begin{bmatrix} Q_1 \\ Q_2 \end{bmatrix}, \tag{25}$$

where Q_1 and Q_2 are respectively $m \times n$ and $(n-m) \times n$ submatrices of Q. The pseudo-inverse (21) can now be efficiently computed according to

$$A_x^+ = Q_1^T L^{-1}, \tag{26}$$

while the columns of Q_2^T form an orthonormal basis of $\mathcal{N}(A_x)$; hence,

$$Z_x = Q_2, \qquad Z_x^- = Q_2^T \tag{27}$$

satisfy the conditions of Proposition 2.1. Notice again that

$$Z_x^{-T} Z_x^- = I.$$

Remark 2.1. The singular value decomposition is commonly used to compute the pseudo-inverse of a matrix, but requires more computational effort than the QR decomposition [see Stewart (Ref. 8)]. It is useful, however, when the matrix is not of full rank, a situation that we have excluded by hypothesis.

We conclude this section by giving an estimate of the coordinate domain $U_x \cap C$ in terms of the map c. Take $U_x \subset B(x, \alpha)$, the open ball in $\mathbb{R}^n$ of centre x and radius $\alpha > 0$. Let

$$\beta = \|A_x^-\|, \qquad \xi = \|Z_x^-\|,$$

$$\gamma_i = \max_{\tilde{x} \in B(x,\alpha)} \max_{\|y\|=1} |c_i''(\tilde{x})(y)(y)|,$$

where $\|\cdot\|$ stands for the l_2 norm; let

$$\gamma = \left(\sum_{i=1}^{n} \gamma_i \right)^{1/2}.$$

Theorem 2.2. The local parametrization θ_x defined by (11) around x maps diffeomorphically the neighborhood $B(0, 1/(2\gamma\xi))$ of the origin in $\mathbb{R}^{n-m}$ onto a neighborhood $U_x \cap C$ of x in C, where $U_x \subset B(x, 1/(\beta\gamma))$.

Proof. The map s_x given by (9) defines, by restriction, a diffeomorphism of $U_x \cap C$ onto a neighborhood V of the origin in $\mathbb{R}^{n-m}$. To get an estimate of U_x, we look for an estimate of V such that, given $z \in V$, there exists a unique $y \in C$ such that $s_x(y) = (0, z)$.

Let

$$\|z\| = \rho.$$

Proposition 2.1 shows that we can look for

$$y = x + Z_x^- z + A_x^- w, \tag{28}$$

with $w \in \mathbb{R}^m$ such that

$$h(w) = c(x + Z_x^- z + A_x^- w) = 0. \tag{29}$$

We shall show that, if

$$\rho \leq 1/(2\beta\gamma\xi),$$

there exists a unique solution of (29) in $B(0, 1/(2\beta^2\gamma))$ by a constructive proof.

Consider the iteration

$$w^{k+1} = w^k - h(w^k), \qquad k = 0, 1, \ldots, \tag{30}$$

starting from $w^\circ = 0 \in \mathbb{R}^m$. We construct a majorizing sequence $\{t_k\}$ of nonnegative real numbers, i.e., such that

$$\|w^k\| \leq t_k, \qquad \text{for all } k.$$

Applying Taylor's formula to the map c around x, we obtain

$$w^{k+1} = -\int_0^1 (1-t)c''(x + tZ_x^- z + tA_x^- w^k)(Z_x^- z + A_x^- w^k)(Z_x^- z + A_x^- w^k)\, dt; \tag{31}$$

define the sequence $\{t_k\}$ by the iteration

$$t_{k+1} = \frac{\gamma}{2}(\beta t_k + \xi\rho)^2, \qquad k = 0, 1, \ldots, \tag{32}$$

starting from $t_0 = 0$. If

$$\|w^k\| \leq t_k,$$

(31) shows that

$$\|w^{k+1}\| \leq t^{k+1},$$

and by induction $\{t_k\}$ is actually a majorizing sequence. It is easy to show that, if

$$\rho \leq 1/2\beta\gamma\xi,$$

the sequence $\{t_k\}$ is monotonically increasing and converges to

$$t_\rho^* = [1-\beta\gamma\xi\rho-(1-2\beta\gamma\xi\rho)^{1/2}]/\beta^2\gamma > 0;$$

hence, w^k remains in $B(0, t_\rho^*)$ for all k.

Consider now $w^{k+1}-w^k$, which can be written as

$$w^{k+1}-w^k = -[h(w^k)-h(w^{k-1})-h'(w^{k-1})(w^k-w^{k-1})]$$
$$-[h'(w^{k-1})-h'(0)](w^k-w^{k-1})+[I-h'(0)](w^k-w^{k-1}).$$

Assuming the induction hypothesis (trivially satisfied for $k=1$)

$$\|w^k-w^{k-1}\| \le t_k - t_{k-1}, \tag{33}$$

we obtain, with the use of the Taylor formula, the estimate

$$\|w^{k+1}-w^k\| \le \frac{\gamma}{2}[\beta^2(t_k-t_{k-1})^2+2\beta^2 t_{k-1}(t_k-t_{k-1})+2\beta\gamma\rho(t_k-t_{k-1})]$$

$$= \frac{\gamma}{2}[(\beta t_k+\xi\rho)^2-(\beta t_{k-1}+\xi\rho)^2] = t_{k+1}-t_k,$$

which proves that (33) holds for all k and that

$$\|w^{k+1}-w^k\| \to 0,$$

since $t_k \nearrow t_\rho^*$; hence, the convergence of the sequence $\{w^k\}$, defined by (30), to w such that $\|w\| \le t_\rho^*$, unique solution of Eq. (29) in $B(0, t_\rho^{**})$, where t_ρ^{**} is the other fixed point of (32) given by

$$t_\rho^{**} = [1-\beta\gamma\xi\rho+(1-2\beta\gamma\xi\rho)^{1/2}]/\beta^2\gamma \ge t_\rho^*.$$

Notice that, for the boundary value $\rho = 1/(2\beta\gamma\xi)$,

$$t_\rho^{**} = t_\rho^* = 1/(2\beta^2\gamma).$$

Then, for any $z \in B(0, 1/(2\beta\gamma\xi))$, there exists a unique solution w of (29) in $B(0, 1/(2\beta^2\gamma))$ and

$$\|y-x\| = \|Z_x^- z + A_x^- w\| \le 1/\beta\gamma. \qquad \square$$

3. Geometry of the Constrained Manifold

3.1. A Riemannian Structure on C. Given a point $x \in C$, we define the *tangent space* T_x to the manifold C at x as the $(n-m)$-dimensional subspace of $\mathbb{R}^n$

$$T_x = \mathcal{N}(A_x) = \{v \in \mathbb{R}^n \mid A_x v = 0\}. \tag{34}$$

If we choose the local coordinate system defined in the neighborhood $U_x \cap C$ of x in C by

$$z(y) = Z_x(y - x), \qquad \text{for } y \in U_x \cap C, \tag{35}$$

we can select the right inverse Z_x^- such that

$$\mathcal{N}(A_x) = \mathcal{R}(Z_x^-);$$

thus,

$$T_x = \{Z_x^- p \,|\, p \in \mathbb{R}^{n-m}\} \in \mathbb{R}^n. \tag{36}$$

The equivalence of (36) with (34) shows that this last definition is independent of the coordinate system [see, e.g., Guillemin and Pollack (Ref. 9)].

It is convenient to endow T_x with a positive bilinear form $\gamma_x : T_x \times T_x \to \mathbb{R}$, called the *Riemannian metric*. The form γ_x is a smooth function of x and defines a *Riemannian structure* γ on C. A natural choice consists in taking

$$\gamma_x^E(v, w) = \langle v, w \rangle_n, \qquad \forall v, w \in T_x \subset \mathbb{R}^n, \tag{37}$$

where $\langle\ ,\ \rangle_n$ denotes the ordinary scalar product on $\mathbb{R}^n$; γ_x^E is called the Riemannian metric on C induced by the Euclidean structure of $\mathbb{R}^n$. In the local coordinate system (35), the induced metric can be expressed as

$$\gamma_x^E(Z_x^- p, Z_x^- q) = \langle p, Z_x^{-T} Z_x^- q \rangle_{n-m}, \qquad \forall p, q \in \mathbb{R}^{n-m},$$

which coincides with the ordinary scalar product on $\mathbb{R}^{n-m}$ iff

$$Z_x^{-T} Z_x^- = I$$

(see Example 2.2 for such a case). It may be convenient to define directly the form in the local coordinate system by

$$\gamma_x^Z(Z_x^- p, Z_x^- q) = \langle p, q \rangle_{n-m}, \qquad \forall p, q \in \mathbb{R}^{n-m}. \tag{38}$$

Notice that γ_x^Z can be viewed as the Riemannian metric on C induced by a scalar product in $\mathbb{R}^n$, defined locally around x by

$$\langle\!\langle v, w \rangle\!\rangle = \langle S_x v, S_x w \rangle_n, \qquad \forall v, w \in \mathbb{R}^n,$$

where S_x is the matrix defined by (12). A general Riemannian structure on C can be defined in a local coordinate system around x by

$$\gamma_x^G(Z_x^- p, Z_x^- q) = \langle p, G_x q \rangle_{n-m}, \tag{39}$$

where

$$G_x = (g_{ij})$$

is a positive-definite symmetric matrix of dimension $n - m$, smooth function

of x. The terminology smooth is used here and in the following sections for continuously differentiable functions to a sufficiently large order for all formulas to have sense.

3.2. Covariant Differentiation and Geodesics. We define a *vector field* V on the submanifold C of $\mathbb{R}^n$ as a smooth map $V: C \to \mathbb{R}^n$ such that $V(x) \in T_x$ for all x.

Let $x \in C$. Given a vector $v \in T_x$ and a vector field W on C, we define a new vector $D_v W \in T_x$, called the *covariant derivative* of W along v; the application $\tau_x(W): T_x \to T_x$, defined by

$$\tau_x(W)(v) = D_v W, \tag{40}$$

must be linear in v and satisfy the chain rule given by

$$\tau_x(fW)(v) = f(x)\tau_x(W)(v) + f'(x)(v)W(x), \tag{41}$$

where f is any real-valued smooth function on C; it specifies an *affine connection* on C at x. Let now V and W be vector fields on C; we define the field $D_V W$, the *covariant derivative* of W with respect to V on C, by its values

$$D_V W(x) = D_v(W), \qquad \text{where } v = V(x) \in T_x. \tag{42}$$

The affine connection is thus specified globally on C [see Milnor (Ref. 10)].

Given the *local coordinate system* (35) around x, the column vectors e_i, $i = 1, \ldots, n-m$, of the matrix Z_x^- form a basis of T_x in $\mathbb{R}^n$. The $\mathscr{C}^{\sigma-1}$ maps $E_i: C \to \mathbb{R}^n$ such that

$$E_i(x) = e_i$$

are vector fields and are said to form the associated *base fields* around x. Properties (41) show that the affine connection on C is determined by the fields $D_{E_i}E_j$. It is customary to express these vector fields in the base fields as

$$D_{E_i}E_j = \sum_{k=1}^{n-m} \Gamma_{ij}^k E_k. \tag{43}$$

The $(n-m)^3$ real-valued smooth functions Γ_{ij}^k determine the connection around x and are called the coefficients of the connection.

A *parametrized curve* in C is a smooth map $x(\cdot)$ from the real numbers into $C \subset \mathbb{R}^n$. The velocity vector field $\dot{x}$ is defined by

$$\dot{x}(t) = dx(t)/dt \in T_{x(t)}, \qquad \text{for all } t \in \mathbb{R}. \tag{44}$$

A vector field V on C defines, by restriction, a vector field $v(\cdot)$ along the

curve $x(\cdot)$ which assigns to each $t \in \mathbb{R}$ a tangent vector

$$v(t) = V(x(t)) \in T_{x(t)}. \tag{45}$$

The vector field $v(\cdot)$ is said to be a *parallel vector field along the curve* $x(\cdot)$, if its covariant derivative with respect to the velocity field $\dot{x}$, denoted Dv/dt, is identically zero, i.e.,

$$Dv/dt = D_{\dot{x}} V = 0. \tag{46}$$

Using the local coordinate system (35) around a point x of the curve and the associated base fields E_i, the vector field $v(\cdot)$ can be uniquely expressed as

$$v = \sum_{i=1}^{n-m} v_i E_i,$$

where v_i are smooth real-valued functions on $\mathbb{R}$, while the velocity field is

$$\dot{x} = \sum_{i=1}^{n-m} (dz_i/dt) E_i,$$

where

$$z_i(t) = z_i[x(t)].$$

It follows from (40), (41), (42), (43), (46) that the functions v_i must satisfy the system of linear differential equations

$$dv_k/dt + \sum_{i,j=1}^{n-m} (dz_i/dt) \Gamma_{ij}^k v_j = 0, \qquad k = 1, 2, \ldots, n-m, \tag{47}$$

and we have the following existence and uniqueness result [see also Hicks (Ref. 11)].

Proposition 3.1. Let $x(\cdot)$ be a parametrized curve in C defined on $[0, T]$. For each vector v in $T_{x(0)}$, there is a unique parallel vector field $v(\cdot)$ along $x(\cdot)$ such that $v(0) = v$. The map $\pi_0^t : T_{x(0)} \to T_{x(t)}$, defined by $\pi_0^t(v(0)) = v(t)$, is a linear isomorphism, called parallel translation along $x(\cdot)$ from $x(0)$ to $x(t)$.

The following result is a fundamental tool in Riemannian geometry [see Milnor (Ref. (10)].

Proposition 3.2. There exists a unique (symmetric) connection on a Riemannian manifold such that parallel translation preserves the Riemannian metric.

In a system of local coordinates $\{z_i\}$, the coefficient functions Γ_{ij}^k defining the connection are uniquely determined by the symmetric matrix function

G, defining the Riemannian metric in (39), and satisfy, for $i, j, k = 1, 2, \ldots, n-m$,

$$\Gamma_{ij}^k = \sum_l \tfrac{1}{2}(\partial g_{jk}/\partial z_i + \partial g_{ik}/\partial z_j - \partial g_{ij}/\partial z_l)(G^{-1})_{lk}, \tag{48}$$

called the second Christoffel identity. Notice that

$$\Gamma_{ij}^k = \Gamma_{ji}^k, \qquad \text{for all } i, j,$$

which indicates that the connection is symmetric. Notice also that, if $G(x)$ remains constant (for instance if $G(x) = I$),

$$\Gamma_{ij}^k \equiv 0.$$

A parametrized curve x of C is called a *geodesic* if its velocity field $\dot{x}$ is parallel along x, i.e.,

$$D\dot{x}/dt = 0.$$

In terms of the local coordinate system (35), the local coordinate functions

$$z = (z_k), \qquad k = 1, \ldots, n-m,$$

must satisfy the system of second-order differential equations

$$d^2 z_k/dt^2 + \sum_{i,j} \Gamma_{ij}^k(z)(dz_i/dt)(dz_j/dt) = 0, \qquad k = 1, \ldots, n-m, \tag{49}$$

derived from (47) with

$$v_k = dz_k/dt.$$

From now on, we consider C endowed with a Riemannian metric and work with the unique connection compatible with it, i.e., the coefficients of which satisfy the second Christoffel identity. The following result about the solutions of (49) establishes the equivalence of our definition of geodesics with the elementary one (curves of minimal lengths).

Proposition 3.3. Let W be a connected compact subset of C in $\mathbb{R}^n$. Given $x \in W$ and $p \in T_x$, there exists a unique geodesic curve $x(\cdot)$ such that

$$x(0) = x, \qquad \dot{x}(0) = p;$$

the map x is either defined for all $t \in \mathbb{R}$ and takes its value in W, or is defined on the interval $[-T, T]$, and $x(T)$ (or $x(-T)$) belongs to the boundary of W. Any two points of W can be joined by a unique geodesic which minimizes the arc length between the points.

See Milnor (Ref. 10) or Bishop and Crittenden (Ref. 12) for a proof. The geodesic curve is denoted by

$$x(t) = \exp_x(tp), \tag{50}$$

and the local coordinate system

$$\exp_x^{-1}: U_x \cap C \to V \subset T_x \simeq \mathbb{R}^{n-m}$$

is called the *normal coordinate system* around x. Observe that, in such coordinates, the geodesic is locally the parametrized line interval

$$\{tp \mid t \in (-\epsilon, +\epsilon)\}.$$

Finally, notice that, if in a local coordinate system the Riemannian metric is defined by a constant metric G, the local coordinate functions

$$z = (z_k), \qquad k = 1, \ldots, n-m,$$

defining the geodesic around x satisfy the second-order differential equation

$$d^2 z_k / dt = 0, \qquad k = 1, 2, \ldots, n-m,$$

since

$$\Gamma_{ij}^k = 0;$$

hence,

$$z(t) = tp,$$

and the local coordinate system is normal.

4. Descent Methods Along Geodesics

4.1. Optimality Conditions. We turn now to the solution of the nonlinear programming problem (1), which we now write

$$\min\{f(x) \mid x \in C\}, \tag{51}$$

i.e., the problem of minimizing the differentiable real-valued function f on the differential manifold C endowed with a smooth Riemannian structure γ.

Given $x \in C$, we define the derivative on C of f at x as the linear form on T_x, $D_C f(x): T_x \subset \mathbb{R}^n \to \mathbb{R}$, such that

$$D_C f(x)(v) = f'(x)(v), \qquad \text{for all } v \in T_x, \tag{52}$$

where $f'(x)$ is the ordinary derivative at x of f considered as a function on $\mathbb{R}^n$.

We can now state a first-order necessary optimality condition for problem (51).

Theorem 4.1. Let $x^* \in C$ be a local minimum of f on the Riemannian manifold C. Then,

$$D_C f(x^*) = 0. \tag{53}$$

Proof. Since x^* is a local minimum of f on C,

$$f(x) \geq f(x^*), \qquad \forall x \in U_{x^*} \cap C, \tag{54}$$

where U_{x^*} is a neighborhood of x^* in $\mathbb{R}^n$. According to Proposition 3.3, (54) is equivalent to

$$f(x(t)) \geq f(x^*), \tag{55}$$

for any geodesic curve $x(\cdot)$ starting from x^* and all $t \in (-\varepsilon, +\varepsilon)$. Applying the mean-value theorem to the function $f \circ x$ between 0 and t, (55) yields

$$f'(x^*)\dot{x}(0) = f'(x^*)(v) \geq 0, \qquad \forall v \in T_{x^*}; \tag{56}$$

hence,

$$D_C f(x^*)(v) = 0, \qquad \forall v \in T_{x^*},$$

by applying (56) to both v and $-v$. □

Conversely, every point $x^* \in C$ such that

$$D_C f(x^*) = 0$$

is called a *critical point* of f.

Given $v \in T_x$, let V be a smooth vector field on C such that

$$V(x) = v.$$

Define the $\mathscr{C}^{\sigma-1}$ differentiable real-valued function $Vf: C \to \mathbb{R}$ by

$$Vf(x) = D_C f(x)(v).$$

Since $\sigma \geq 2$, this function has a derivative on C at x, $D_C Vf(x) \in L(T_x, \mathbb{R})$. If x^* is a *critical point*, we define the Hessian of f on C as the bilinear form $Hf(x^*): T_{x^*} \times T_{x^*} \to \mathbb{R}$ given by

$$Hf(x^*)(v, w) = D_C Vf(x^*)(w), \qquad \forall v, w \in T_{x^*}. \tag{57}$$

This form is well defined (i.e., independent of the choice of the vector field V) and symmetric [see Milnor (Ref. 10)]. This definition holds only at a critical point.

A *critical point* x^* is said to be *nondegenerate* iff $Hf(x^*)$ is nondegenerate, i.e.,

$$Hf(x^*)(v, v) = 0 \qquad \text{implies } v = 0.$$

In the following, we assume that all the critical points of f on C are nondegenerated; f is called a *Morse function* on C [see Milnor (Ref. 10)]. It follows that the critical points are isolated. We can then give a second-order optimality condition in stronger terms than usually expressed [see, e.g., Luenberger (Ref. 13)].

Theorem 4.2. A point $x^* \in C$ is a strict local minimum of a Morse function f on the Riemannian manifold C iff x^* is a critical point of f and the Hessian form $Hf(x^*)$ is positive definite.

Proof. Let $x(\cdot)$ be the geodesic curve starting from x^* with tangent v; and let V be the unique parallel vector field along $x(\cdot)$, such that

$$V(x^*) = v.$$

Applying the Taylor formula to the $\mathscr{C}^\sigma$ differentiable function $f \circ x : \mathbb{R} \to \mathbb{R}$ yields

$$f(x(t)) - f(x^*) = D_C f(x^*)(v)t + \tfrac{1}{2} D_C V f(x(\theta t))(V(x(t))t^2, \qquad (58)$$

with $\theta \in (0, 1)$. If x^* is a local minimum, the first member on the right-hand side vanishes by Theorem 4.1, while the second member must be nonnegative; by continuity,

$$Hf(x^*)(v, v) \geq 0, \qquad \forall v \in T_{x^*};$$

and, since f is a Morse function, strict inequality holds for all $v \neq 0$, and x is a strict local minimum. Conversely, if x^* is a critical point such that $Hf(x^*)$ is positive definite, the left-hand side of (58) remains strictly positive for all $t \neq 0$ sufficiently small and all $v \neq 0 \in T_{x^*}$, which shows that x^* is a strict local minimum.

Remark 4.1. *Generic Properties.* Morse functions form an open dense subset of the space of $\mathscr{C}^\sigma$ real-valued functions on C for $2 \leq \sigma \leq +\infty$ [see Hirsch (Ref. 14) for a precise definition of the topology]. In other words, by a *slight perturbation*, it is always possible to ensure that f is a Morse function. Such a property is said to be generic [see Golubitsky and Guillemin (Ref. 15)].

Incidentally, the other assumption that we made (namely, 0 is a regular value of the $\mathscr{C}^0$ map $C : \mathbb{R}^n \to \mathbb{R}^m$) is also generic for $1 \leq \sigma \leq +\infty$. In terms of differential topology, the assumption amounts to saying that c is trans-

verse to 0 on the manifold $C(c \pitchfork \{0\})$. The genericity of the assumption follows from the transversality theorem [see, e.g., Ref. 14].

These remarks show that our assumptions hold for most $\mathscr{C}^\sigma$ problems (1).

Remark 4.2. The proofs of Theorem 4.1 and 4.2 use explicitly the Riemannian structure of C, since they introduce geodesics. They generalize similar results for the unconstrained case, which actually make use of the Euclidean structure of $\mathbb{R}^n$. Such results hold, however, on a manifold without a Riemannian structure and can be established using a local coordinate system and showing that they are independent of its choice.

Remark 4.3. The reader may at this point wonder how these results relate to the classical formulation of the optimality conditions for problem (1) in terms of Lagrange multipliers. The first-order optimality condition (53) is equivalent to

$$f'(x^*)(v) = \langle \nabla f(x^*), v \rangle_n = 0, \qquad \forall v \in T_{x^*}, \tag{59}$$

by definition of the gradient ∇f of f in the Euclidean space $\mathbb{R}^n$; condition (59) expresses that $\nabla f(x^*)$ is orthogonal to T_{x^*}, hence belongs to $\mathscr{R}(A_{x^*}^T)$. Thus, there exists $\lambda^* \in \mathbb{R}^m$ such that

$$\nabla f(x^*) + \nabla c(x^*)\lambda^* = 0, \tag{60}$$

the classical form for the first-order optimality condition in terms of the Lagrangian

$$l(x, \lambda) = f(x) + \langle \lambda, c(x) \rangle_m. \tag{61}$$

Given $x \in C$, let

$$\lambda(x) = -(\nabla c(x)^T \nabla c(x))^{-1} \nabla c(x)^T \nabla f(x); \tag{62}$$

clearly, $\nabla_x l(x, \lambda(x)) \in T_x$, and $\lambda^* = \lambda(x^*)$ satisfies (60). At the critical point x^*,

$$Hf(x^*)(v, v) = \nabla_{xx}^2 l(x^*, \lambda^*)(v, v), \qquad \forall v \in T_{x^*}, \tag{63}$$

and the second-order optimality condition of Theorem 4.2 can be stated in terms of the positive-definiteness of the restriction to T_{x^*} of the Hessian of the Lagrangian $\nabla_{xx}^2 l(x^*, \lambda^*)$. See Luenberger (Ref. 13).

4.2. Gradient of a Function on C. Let γ be the Riemannian structure on C. We define the gradient on C of f at x as the tangent vector $\nabla_C^\gamma f(x) \in T_x$ such that

$$\gamma_x(\nabla_C^\gamma f(x), v) = D_C f(x)(v), \qquad \text{for all } v \in T_x. \tag{64}$$

The vector $\nabla^{\gamma}_{C} f(x)$ clearly depends upon the Riemannian metric. Notice that $\nabla^{\gamma}_{C} f$ is a vector field on C, called the *gradient field.*

In the local coordinate system around x,

$$z(y) = Z_x(y - x), \qquad \text{for all } y \in U_x \cap C, \tag{65}$$

there are two natural Riemannian metrics as discussed in Section 3.1. If we use the metric γ^{Z}_{x} defined in (38) as the scalar product in $\mathbb{R}^{n-m}$, we obtain

$$\nabla^{Z}_{C} f(x) = Z^{-}_{x} g^{Z}_{x}, \qquad \text{with } g^{Z}_{x} = Z^{-T}_{x} \nabla f(x) \in \mathbb{R}^{n-m}; \tag{66}$$

while, if we use the metric γ^{E}_{x} induced by the scalar product in $\mathbb{R}^{n}$ defined in (37), we obtain

$$\nabla^{E}_{C} f(x) = Z^{-}_{x} g^{E}_{x}, \qquad \text{with } g^{E}_{x} = (Z^{-T}_{x} Z^{-}_{x})^{-1} g^{Z}_{x}. \tag{67}$$

We call g^{Z}_{x} and g^{E}_{x}, respectively, the *reduced gradient* and the *Euclidean reduced gradient*; they coincide obviously iff

$$Z^{-T}_{x} Z^{-}_{x} = I,$$

which gives an argument in favor of the coordinate system presented in Example 2.2. The terminology *reduced gradient* introduced by Abadie (Ref. 5) expresses the reduction of dimension achieved by considering vectors in $\mathbb{R}^{n-m}$ and is employed here in a more general sense than in Gabay and Luenberger (Ref. 1), where only the partitioned coordinate system of Example 2.1 was considered. Notice that

$$\nabla^{E}_{C} f(x) = P^{E}_{x} \nabla f(x), \tag{68}$$

where P^{E}_{x} is the orthogonal projector onto T_x in the Euclidean space $\mathbb{R}^{n}$; $\nabla^{E}_{C} f(x)$ will be called the (Euclidean) *projected gradient*, following Rosen (Ref. 3). The vector $\nabla^{Z}_{C} f(x)$ can also be interpreted as the orthogonal projection of ∇f onto T_x, but with respect to the scalar product

$$\langle\!\langle v, w \rangle\!\rangle = \langle S_x v, S_x w \rangle_n, \tag{69}$$

with S_x defined by (12).

Remark 4.4. Another argument in favor of the local coordinate system defined in Example 2.2 arises from the relationship between the projected gradient $\nabla^{E}_{C} f(x)$ and the gradient $\nabla_x l(x, \lambda(x))$ of the Lagrangian. Both vectors are in T_x, provided $\lambda(x)$ is given by (62), which can be expressed (and computed) in terms of the pseudo-inverse (26) of $A_x = \nabla c(x)^{T}$,

$$\lambda(x) = -A^{+T}_{x} \nabla f(x). \tag{70}$$

We then have

$$\nabla_C^E f(x) = P_x^E \nabla f(x) = \nabla_x l(x, \lambda(x)), \tag{71}$$

a formula used by Luenberger (Ref. 2) to define the gradient of f under constraints [see also Hestenes (Ref. 16)].

If we use a general right inverse A_x^- and define

$$\mu(x) = -A_x^{-T} \nabla f(x), \tag{72}$$

then by (17)

$$\nabla_x l(x, \mu(x)) = P_x^T \nabla f(x) = Z_x^T Z_x^{-T} \nabla f(x) = Z_x^T g_x^Z, \tag{73}$$

which differs from $\nabla_C^Z f(x)$ and actually does not belong to T_x, except if

$$Z_x^- = Z_x^T,$$

i.e., if the columns of Z_x^T are orthonormal.

4.3. Steepest Descent Method along Geodesics. To find a local minimum on $\mathbb{R}^n$ of a continuously differentiable function f, the steepest descent method generates, starting from an initial estimate x°, successive approximations according to the iteration

$$x^{k+1} = x^k + t_k p^k, \qquad k = 0, 1, \ldots, \tag{74}$$

where p^k is the direction which minimizes

$$\langle \nabla f(x^k), p \rangle_n / \|p\|_n,$$

namely,

$$p^k = -\nabla f(x^k), \tag{75}$$

called the *direction of steepest descent*; and the *stepsize* t_k is a positive scalar selected, for instance, as the first local minimum on $\mathbb{R}^+$ of the function

$$j(t) = f(x^k + tp^k);$$

we say that t_k is determined by an *exact line search* and denote the solution by

$$t_k = \arg\min\{f(x^k + tp^k) \mid t \geq 0\}. \tag{76}$$

This method can be generalized to find a local minimum of f on a *Riemannian manifold* C. The direction of steepest descent for f on C at x^k is given by

$$p^k = -\nabla_C^\gamma f(x^k), \tag{77}$$

which minimizes

$$\gamma_{x^k}(\nabla^{\gamma}_{C} f(x^k), p)/\|p\|_{\gamma_{x^k}}, \qquad \text{for all } p \in T_x^k, \|p\|_{\gamma_{x^k}} = (\gamma_{x^k}(p,p))^{1/2}.$$

Proposition 3.3 shows that the geodesics of C play the role of the straight lines in $\mathbb{R}^n$; we thus define the *steepest descent method along geodesics* as the iteration

$$x^{k+1} = \exp_{x^k}(t_k p^k), \tag{78}$$

where p^k is defined by (77) and the stepsize t_k is determined by an *exact geodesic search*,

$$t_k = \arg\min\{f(\exp_{x^k}(tp^k)) \mid t \in \mathbb{R}^+\}. \tag{79}$$

Algorithm (77), (78), (79) has been first introduced and analyzed by Luenberger (Ref. 2), who explicitly used the Riemannian structure γ^E on C induced by the Euclidean structure of $\mathbb{R}^n$. As noticed in (68), $\nabla^E_C f(x^k)$ is then the orthogonal projection of the gradient $\nabla f(x^k)$ on T_{x^k}; hence, the terminology of gradient projection method along geodesics. In his paper, Luenberger established the global convergence of the algorithm to a critical point of f on C and estimated the speed of convergence in the neighborhood of a critical point which is a strict local minimum. Lichnewsky (Ref. 4) has recently proposed the algorithm (77), (78), (79) for a general Riemannian manifold and established similar results; he also studied a conjugate gradient version of it.

We give a global convergence theorem which generalizes the classical results for the method in $\mathbb{R}^n$ [see Polak (Ref. 17) and Ortega and Rheinboldt (Ref. 18)]. We first need to introduce the following notation: let W_k denote the connected component containing x^k of the level set

$$\{x \in C \mid f(x) \leq f(x^k)\}.$$

Theorem 4.3. Assume that f is continuously differentiable and that W_0 is compact. Then the sequence $\{x^k\}$ constructed by the steepest descent method along geodesics (77), (78), (79) is well defined; it is either finite, terminating at a critical point, or is infinite and every accumulation point is a critical point. If the critical values of f are distinct, the whole sequence $\{x^k\}$ converges to a critical point.

Proof. The compactness of W_0 implies the compactness of the closed subsets W_k, since the sequence $\{f(x^k)\}$ is monotone and nonincreasing. If x^k is a critical point, $p^k = 0$ and the algorithm does not generate new approximations; introducing a stopping test, it can be terminated at iteration

k. Assume now that x^k is not a critical point; hence,

$$\gamma_{x^k}(\nabla^{\gamma}_{C} f(x^k), p^k) = -\gamma_{x^k}(\nabla^{\gamma}_{C} f(x^k), \nabla^{\gamma}_{C} f(x^k)) < 0. \tag{80}$$

Denote the arc of geodesic starting from x^k with tangent p^k by

$$x(t) = \exp_{x^k}(tp^k), \tag{81}$$

and introduce the function $j: \mathbb{R}^+ \to \mathbb{R}$ defined by

$$j(t) = f[x(t)]. \tag{82}$$

Let

$$\bar{t} = \limsup J,$$

where the set J is defined by

$$J = \{t > 0 \,|\, x(t) \text{ is defined and } j(t) < j(0) = f(x^k)\}; \tag{83}$$

by (80), the set J is nonempty; and, since W_k is compact, Proposition 3.3 implies that either $\bar{t} = +\infty$, and $x(t) \in W_k$, for all $t \in [0, +\infty)$, or $\bar{t}$ is finite, $f(x(\bar{t})) = f(x^k)$, and $x(t) \in W_k$, for all $t \in [0, \bar{t}]$. In both cases, the stepsize rule (79) is well defined,

$$t_k \in (0, \bar{t}).$$

Observe that t_k satisfies

$$j'(t_k) = \gamma_{x^{k+1}}(\nabla^{\gamma}_{C} f(x^{k+1}), \pi_0^{t_k}(p^k)) = \gamma_{x^k}(\pi_{t_k}^{\circ}(\nabla^{\gamma}_{C} f(x^{k+1})), p^k) = 0, \tag{84}$$

where $\pi_0^{t_k}(p^k)$ is the vector of $T_{x^{k+1}}$ obtained from p^k by parallel translation along the curve $x(\cdot)$ from x^k to x^{k+1} (see Proposition 3.1); the second equality of (84) results from the definition of the geodesic $x(\cdot)$. Observe also, that given $\alpha \in (0, \frac{1}{2})$, the equation

$$\gamma_{x^k}(\pi_t^{\circ}(\nabla^{\gamma}_{C} f(x(t))), p^k) = \alpha \gamma_{x^k}(\nabla^{\gamma}_{C} f(z^k), p^k) \tag{85}$$

has a smallest solution $\hat{t} \in (0, t_k)$ and

$$\gamma_{x^k}(\pi_t^{\circ}(\nabla^{\gamma}_{C} f(x(t))), p^k) < \alpha \gamma_{x^k}(\nabla^{\gamma}_{C} f(x^k), p^k), \qquad \text{for all } t \in [0, \hat{t}).$$

Applying the mean-value theorem, we obtain the estimate

$$f(x^{k+1}) - f(x^k) < f(x(\hat{t})) - f(x^k) < -\alpha \hat{t} \|\nabla^{\gamma}_{C} f(x^k)\|^2_{\gamma_{x^k}}.$$

The sequence $\{f(x^k)\}$ is monotone decreasing and is bounded from below, since the continuous function f attains its minimum on the compact W_0; hence, it converges to a limit. Let $\{x^{k_i}\}$ be a subsequence of $\{x^k\}$ converging to $x^* \in W_0$. Suppose that x^* is not a critical point, i.e.,

$$\|\nabla^{\gamma}_{C} f(x^*)\|_{\gamma_{x^*}} = \delta > 0.$$

The continuity of the Riemannian metric γ and of the gradient field $\nabla^{\gamma}_{C} f$ implies that

$$\|\nabla^{\gamma}_{C} f(x^{k_i})\|_{\gamma_{x^{k_i}}} \geq \delta/2, \qquad \text{for all } i > I;$$

hence,

$$f(x^{k_{i+1}}) < f(x^{k_i+1}) < f(x^{k_i}) - \alpha \hat{t} \delta^2/4,$$

which contradicts the fact that $\{f(x^{k_i})\}$ converges to $f(x^*)$. Thus, x^* is a critical point.

Finally, suppose that x^* and x^{**} are distinct accumulation points of the sequence $\{x^k\}$ in W_0: x^* and x^{**} are critical points of f. Since $\{f(x^k)\}$ converges, we must have

$$f(x^*) = f(x^{**}),$$

which is impossible if the critical values of f are distinct.

Remark 4.5. The proof shows that Theorem 4.3 holds if we replace the stepsize rule (79) by rule (84) or rule (85) or by the following rule: find

$$t_k = 2^{-l} \hat{t},$$

where $\hat{t}$ is an initial guess and l is the smallest integer satisfying, for a given $\alpha \in (0, \frac{1}{2})$,

$$f(x(t_k)) \leq f(x^k) - \alpha t_k \|\nabla^{\gamma}_{C} f(x^k)\|^2_{\gamma_{x^k}}. \tag{86}$$

By analogy with stepsize selection rules for unconstrained minimization, we call (84), (85), and (86), respectively the Curry, Altman, and Armijo principles [see Ortega and Rheinhold (Ref. 18)].

Remark 4.6. Notice that the algorithm may converge to a critical point which is not a local minimum but a saddle point. It is useful to diagnose when the algorithm reaches a neighborhood of such a point and to generate a new direction of descent away from it. Methods have been proposed recently for the unconstrained case (Refs. 19, 20); however, they require second-order information. In Ref. 4, a particular scheme is proposed for the constrained case, and the modified algorithm is shown to converge to a local minimum if f is a $\mathscr{C}^2$ Morse function and has distinct critical values.

We give finally an estimate of the speed of convergence.

Theorem 4.4. Assume that f is a $\mathscr{C}^{\sigma}$ Morse function on the Riemannian manifold C, with $\sigma \geq 3$. Suppose that the steepest descent method along geodesics (77), (78), (79) generates a sequence $\{x^k\}$ converging to a

critical point x^* in C such that the Hessian form $Hf(x^*)$ satisfies

$$m\|v\|_{\gamma^*}^2 \leq Hf(x^*)(v, v) \leq M\|v\|_{\gamma^*}^2, \qquad \text{for all } v \in T_{x^*}, \tag{87}$$

where m and M are two positive scalars. Then, the sequence $\{x^k\}$ is linearly convergent and

$$\lim_{k \to +\infty} \left[\frac{f(x^{k+1}) - f(x^*)}{f(x^k) - f(x^*)}\right] \leq \left[\frac{M-m}{M+m}\right]^2. \tag{88}$$

See Lichnewsky (Ref. 4) for a proof [or Luenberger (Ref. 2) for the special case of the Riemannian metric induced by the Euclidean structure of $\mathbb{R}^n$]. This result extends the estimate of rate of convergence for the steepest descent method in $\mathbb{R}^n$ [see Luenberger (Ref. 13)]. See also Gabay and Luenberger (Ref. 1), where this estimate was obtained for the special normal coordinate system described in Example 2.1.

Notice that the steepest descent method depends upon the Riemannian metric on C as well as the scalars m, M defined by (87) and giving the estimate of the speed of convergence. It is therefore important to use information about the scaling of the variables in order to improve the conditioning of the problem and obtain better computational efficiency. The framework that we have developed shows that such information should be embodied in the Riemannian structure, i.e., in defining the scalar product in the tangent space.

4.4. Newton's Method along Geodesics. Let x^* be a local unconstrained minimum of a $\mathscr{C}^\sigma$ function f, $\sigma \geq 3$, such that the Hessian form $\nabla^2 f(x^*)$ is positive definite; and let U_* be a neighborhood of x^* such that, for all $x \in U_*$, $\nabla^2 f(x)$ is positive definite. Starting from $x^\circ \in U_*$, Newton's method for unconstrained minimization generates a sequence of approximations of x^* according to the iteration

$$x^{k+1} = x^k - (F_k)^{-1} \nabla f(x^k),$$

where F_k denotes the (nonsingular) symmetric matrix defining the Hessian form $\nabla^2 f(x^k)$; provided U_* is sufficiently small, the iterates remain in U_*, and the sequence $\{x^k\}$ is well defined and converges to x^* with a quadratic rate of convergence [see, e.g., Ortega and Rheinboldt (Ref. 18)].

The extension of Newton's method to the minimization on a manifold C presents a major difficulty, since it is not possible to define the Hessian form of f on C outside of a critical point (see Section 4.1). However, we can define, at a noncritical point $x \in C$, a quadratic form on the tangent space T_x by exploiting the Riemannian structure of the manifold C.

Let $v \in T_x$, and consider the geodesic curve $x(\cdot)$ starting from x and tangent to v,

$$x(t) = \exp_x(tv). \tag{89}$$

We proceed as for the definition of the Hessian, but let now V be the *parallel vector field* along the curve $x(\cdot)$ such that $V(0) = v$; it is unique, by Proposition 3.1. Define the $\mathscr{C}^\sigma$ function $j: \mathbb{R}^+ \to \mathbb{R}$ by

$$j(t) = f(x(t)). \tag{90}$$

By definition of the gradient of f on C with respect to the Riemannian metric γ, we have

$$j'(t) = \gamma_{x(t)}(\nabla_C f(x(t)), V(t)) \tag{91}$$

and

$$\begin{aligned} j''(t) &= \gamma_{x(t)}\left(\frac{D\nabla_C f}{dt}(x(t)), V(t)\right) + \gamma_{x(t)}\left(\nabla_C f(x(t)), \frac{DV}{dt}(t)\right) \\ &= \gamma_{x(t)}\left(\frac{D\nabla_C f}{dt}(x(t)), V(t)\right), \end{aligned} \tag{92}$$

since V is a parallel vector field. We define $F(x): T_x \times T_x \to \mathbb{R}$ by

$$F(x)(v, v) = j''(0) = \gamma_x(D_v \nabla_C f, v), \qquad \text{for all } v \in T_x. \tag{93}$$

The covariant derivative $D_v \nabla_C f$ of the gradient field $\nabla_C f$ along the vector v is linear in v [see (40)]; hence, $F(x)$ is a *quadratic form.* The regularity of the solution (89) of the differential equations (48) defining the geodesic with respect to the initial conditions implies that F is $\mathscr{C}^{\sigma-2}$ differentiable with respect to x. Notice that, at a critical point x^*, $F(x^*)$ coincides with the Hessian form defined by (57):

$$F(x^*)(v, v) = Hf(x^*)(v, v), \qquad \text{for all } v \in T_{x^*}. \tag{94}$$

Define the map $F_x: T_x \to T_x$ by

$$F(x)(v, v) = \gamma_x(F_x v, v), \qquad \text{for all } v \in T_x; \tag{95}$$

notice that F_x is simply the Riemannian connection of the gradient field at x,

$$F_x = \tau_x(\nabla_C f). \tag{96}$$

It is a linear isomorphism of T_x and is self-adjoint (with respect to the Riemannian metric γ_x); denote by $(F_x)^{-1}$ its inverse (defined on T_x).

Let x^* be a nondegenerated local minimum of f on C; by Theorem 4.2, the Hessian $Hf(x^*)$ is positive definite. Identity (94), together with

the continuity of F, show that there exists a neighborhood $U_* \cap C$ of x^* in C such that $F(x)$ is positive definite for all $x \in U_* \cap C$. Starting from $x^\circ \in U_* \cap C$, *Newton's method along geodesics* generates the sequence of successive approximations $\{x^k\}$ according to the iteration

$$x^{k+1} = \exp_{x^k}(p^k), \tag{97}$$

where the *Newton's direction* $p^k \in T_k$ is given by

$$p^k = -(F_k)^{-1} \nabla_C f(x^k). \tag{98}$$

The method is well defined provided $U_* \cap C$ is chosen small enough so that the sequence $\{x^k\}$ remains in this neighborhood, and $\{x^k\}$ converges to x^* quadratically.

Theorem 4.5. Assume that f is a $\mathscr{C}^\sigma$ Morse function on the Riemannian manifold C, with $\sigma \geq 3$. Let x^* be a local minimum of f on C. Then, there exists a neighborhood $U_* \cap C$ of x^* in C such that, if $x^\circ \in U_* \cap C$, Newton's method along geodesics generates a sequence $\{x^k\}$ in $U_* \cap C$ converging to x^*, and there exists a constant K such that

$$\delta(x^{k+1}, x^*) \leq K(\delta(x^k, x^*))^2, \qquad \text{for all } k = 0, 1, \ldots, \tag{99}$$

where $\delta(\cdot, \cdot)$ stands for the Riemannian distance.

Proof. Given $x \in C$ and $v \in T_x$, define the $\mathscr{C}^{\sigma-1}$ function $\varphi: \mathbb{R}^+ \to \mathbb{R}$ by

$$\varphi(t) = \gamma_{x(t)}(\nabla_C f(x(t)), W(t)), \tag{100}$$

where $W(t)$ is a parallel vector field along the curve $x(\cdot)$ given by (89). Notice that

$$\varphi'(t) = \gamma_{x(t)}(F_{x(t)} V(t), W(t)), \tag{101}$$

where $V(t)$ is the parallel vector field along $x(\cdot)$ such that

$$V(0) = v.$$

Consider first the geodesic curve $x(\cdot)$ starting at x^k and tangent to p^k given by (97). Taylor's formula for the function $\varphi(t)$,

$$\varphi(1) = \varphi(0) + \varphi'(0) + \int_0^1 [\varphi'(t) - \varphi'(0)]\, dt,$$

yields

$$\gamma_{x^{k+1}}(\nabla_C f(x^{k+1}), W(1)) = \int_0^1 \gamma_{x^{k+1}}(\pi_t^1 [F_{x(t)} - \pi_0^t F_{x^k} \pi_t^0] V(t), W(1))\, dt; \tag{102}$$

since F_x is $\mathscr{C}^{\sigma-2}$ differentiable on $U_* \cap C$, we obtain the estimate

$$\|\nabla_C f(x^{k+1})\|_\gamma \leq (L/2)\|p^k\|_\gamma^2 \leq (L/2m^2)\|\nabla_C f(x^k)\|_\gamma^2, \tag{103}$$

if

$$\|(F_x)^{-1}\| \leq 1/m, \qquad \text{for all } x \in U_* \cap C.$$

Consider now the minimal geodesic joining x^{k+1} and x^*, which exists by Proposition 3.3; we can find $q^{k+1} \in T_{k+1}$ such that the curve

$$x(t) = \exp_{x^{k+1}}(tq^{k+1})$$

satisfies

$$x(0) = x^{k+1} \quad \text{and} \quad x(1) = x^*.$$

For this choice of $x(\cdot)$, Taylor's formula for the function φ yields

$$\gamma_{x^*}(\nabla_c f(x^*), W(1)) = 0 = \gamma_{k+1}(\nabla_C f(x^{k+1}), W(0)) + \int_0^1 \gamma_{x(t)}(F_{x(t)}Q(t), W(t))\, dt; \tag{104}$$

assuming that

$$m\|v\|_{\gamma_x}^2 \leq \gamma_x(F_x v, v) \leq M\|v\|_{\gamma_x}^2, \qquad \text{for all } x \in U_* \cap C,\ v \in T_x,$$

we obtain the estimate

$$m\|q^{k+1}\|_\gamma \leq \|\nabla_C f(x^{k+1})\|_\gamma \leq M\|q^{k+1}\|_\gamma. \tag{105}$$

Combining (103) and (105) for k and $k-1$, and noticing that

$$\delta(x^{k+1}, x^*) = \|q^{k+1}\|_\gamma,$$

we obtain

$$\delta(x^{k+1}, x^*) \leq (LM/2m^3)(\delta(x^k, x^*))^2, \tag{106}$$

which shows the quadratic convergence of $\{x^k\}$ to x^*, provided

$$\delta(x^0, x^*) \leq 2m^3/LM.$$

4.5. Quasi-Newton Methods along Geodesics. It is possible to modify Newton's method along geodesics in order to have a globally convergent method, i.e., a method which constructs a sequence $\{x^k\}$ converging to a critical point x^* from any initial approximation $x^\circ \in C$. The modified method consists in generating the sequence $\{x^k\}$ according to the iteration

$$x^{k+1} = \exp_{x^k}(t_k p^k), \tag{107}$$

where p^k is a descent direction defined by

$$p^k = -(G_k)^{-1}\nabla_C f(x^k), \tag{108}$$

and the stepsize t_k is chosen according to the Armijo principle (86); i.e., given $\alpha \in (0, \frac{1}{2})$, let

$$t_k = 2^{-l},$$

with l the smallest integer such that

$$f(\exp_{x^k}(2^{-l}p^k)) \leq f(x^k) - \alpha 2^{-l}\gamma_{x^k}(\nabla_C f(x^k), (G_k)^{-1}\nabla_C f(x^k)). \tag{109}$$

We now specify how the operator G_k on T_k must be chosen in order that p^k approximates the Newton direction (98). Observe that (102) yields, for Newton's method, the equation

$$\nabla_C f(x^{k+1}) = \pi_0^1(\nabla_C f(x^k)) + \int_0^1 \pi_t^1 F_{x(t)} \pi_0^t(p^k)\, dt. \tag{110}$$

Define the linear operator $\bar{F}_{k+1}: T_{k+1} \to T_{k+1}$ by

$$\bar{F}_{k+1} = \int_0^1 \pi_t^1 F_{x(t)} \pi_1^t\, dt. \tag{111}$$

Equation (110) can be rewritten as

$$\nabla_C f(x^{k+1}) - \pi_0^1(\nabla_C f(x^k)) = \bar{F}_{k+1}\pi_0^1(p^k); \tag{112}$$

it is thus natural, in order to approach Newton's method, to require that the operator G_{k+1} satisfies the *quasi-Newton equation for constrained minimization*

$$\nabla_C f(x^{k+1}) - \pi_0^{t_k}\nabla_C f(x^k) = G_{k+1}\pi_0^{t_k}(t_k p^k), \tag{113}$$

where we have slightly modified (112) to take into account the presence of the stepsize t_k [we recall that $\pi_0^{t_k}$ denote the parallel translation along the curve $x(t) = \exp_{x^k}(tp^k)$ from $x(0) = x^k$ to $x(t_k) = x^{k+1}$]. Many updating formula for G_k can be proposed which satisfy the quasi-Newton equation and generalize to the constrained case the constellation of quasi-Newton methods for the minimization of f on $\mathbb{R}^n$ [see the excellent survey by Dennis and Moré (Ref. 21)]. We can specialize further the choice of the updating scheme by requiring some additional properties. For instance, we want that p^k be always a descent direction in order to find an admissible stepsize t_k satisfying (109); hence, the updating formulae should generate, from a positive-definite operator G_k on T_k, a new operator G_{k+1} which is positive definite on T_{k+1}. In order to approximate properly the self-adjoint operator $(F_k)^{-1}$, we can insist on having self-adjoint operators $(G_k)^{-1}$; hence, the

updating formula should preserve the symmetry of the operators $(G_k)^{-1}$. These two requirements, in addition to the quasi-Newton equation (113), are satisfied by a family of rank-two updating formulas. Because of the recognized superiority of the Broyden–Fletcher–Goldfarb–Shanno updating scheme for minimization in $\mathbb{R}^n$, we establish the corresponding formula for the constrained case. Denoting

$$s^k = \pi_0^{t_k}(t_k p^k) \in T_{k+1}, \tag{114}$$

$$y^k = \nabla_C f(x^{k+1}) - \pi_0^{t_k} \nabla_C f(x^k) \in T_{k+1}, \tag{115}$$

we define the operator $G_{k+1}: T_{k+1} \to T_{k+1}$ by

$$G_{k+1}p = \tilde{G}_k p - \frac{\gamma_{k+1}(s^k, \tilde{G}_k p)}{\gamma_{k+1}(s^k, \tilde{G}_k s^k)} \tilde{G}_k s^k + \frac{\gamma_{k+1}(y^k, p)}{\gamma_{k+1}(y^k, s^k)} y^k, \qquad \text{for all } p \in T_{k+1}, \tag{116}$$

with

$$\tilde{G}_k = \pi_0^{t_k} G_k \pi_{t_k}^0.$$

If G_k is positive definite on T_k, G_{k+1} is positive definite on T_{k+1} iff

$$\gamma_{k+1}(y^k, s^k) > 0; \tag{117}$$

we call (116) the *generalized BFGS update formula* for constrained minimization.

Notice that the computation of G_{k+1} requires only first-order information, namely the gradient at x^k and x^{k+1}, a definite advantage over the operator F_k used in Newton's method, which involves second-order information. To prove the global convergence of the method (107), (108), (109), with update (116), we must extend the very technical analysis of the BFGS method by Powell (Ref. 22) to the constrained case. This can be achieved, provided there exists a constant M such that the inequality

$$\gamma_{k+1}(y^k, y^k) \leq M \gamma_{k+1}(y^k, s^k) \tag{118}$$

holds for all k; we must also observe that the stepsize rule (109) implies

$$\gamma_{k+1}(\nabla_C f(x^{k+1}), \pi_0^{t_k} p^k) \geq \alpha' \gamma_k(\nabla_C f(x^k), p^k), \tag{119}$$

with $0 < \alpha < \alpha' < 1$. We can then establish a result similar to the one of Powell.

Theorem 4.6. Assume that f is a $\mathscr{C}^\sigma$ Morse function (with distinct critical values) on the Riemannian manifold C, with $\sigma \geq 2$, and that the

level set W_0 is compact. Let G_0 be any positive-definite, self-adjoint operator on T_0. Suppose that the sequence $\{x^k\}$ constructed by the quasi-Newton method along geodesics (107), (108), (109), with update (116), satisfies (118). Then, the sequence $\{x^k\}$ either terminates at or converges to a critical point.

The superiority of this quasi-Newton method along geodesics over the steepest descent method along geodesics of Section 4.3 is evidenced by the following result on its speed of convergence.

Theorem 4.7. Assume that f is a $\mathscr{C}^\sigma$ Morse function on the Riemannian manifold C, with $\sigma \geq 2$. If the sequence $\{x^k\}$ constructed by the quasi-Newton method along geodesics converges to a critical point x^* such that the Hessian form $Hf(x^*)$ is positive definite, then the sequence $\{x^k\}$ is superlinearly convergent, i.e.,

$$\lim_{k \to +\infty} \left[\frac{\delta(x^{k+1}, x^*)}{\delta(x^k, x^*)} \right] = 0. \tag{120}$$

Proof. We simply outline the argument. Extending to the constrained case Powell's technique (Ref. 22), we first show that

$$\sum_{k=0}^{+\infty} \delta(x^k, x^*)$$

is bounded. We then use estimates of the type of Dennis and Moré (Ref. 23) to show that

$$\lim_{k \to +\infty} \left[\frac{\| [G_k - F_k] p^k \|_\gamma}{\| p^k \|_\gamma} \right] = 0. \tag{121}$$

This implies that, after a finite number of iterations, the stepsize

$$t_k = 1$$

satisfies the test (109), provided $\alpha < \frac{1}{2}$ and the limit (120) holds.

Remark 4.7. Condition (121) expresses that the direction p^k given by (108) approaches asymptotically the Newton direction (98).

Remark 4.8. We could define similarly a generalized Davidon–Fletcher–Powell update; but the stepsize t_k of the corresponding quasi-Newton method along geodesics must then be chosen by an exact geodesic search (79) in order to show the global and superlinear convergence.

5. Practical Implementation

5.1. Coordinate System. In practice, the manifold C must be described by an atlas of local coordinate systems. Let x^k be the current of any of the methods presented in Section 4; and let A_k be the $m \times n$ Jacobian matrix of the map c defining C. In the neighborhood $U_k \cap C$ of x^k in C, we use the local coordinate system

$$z_k(x) = Z_k(x - x^k), \tag{122}$$

where Z_k is an $(n-m) \times n$ matrix of right inverse Z_k^-, both chosen in terms of A_k and its right inverse A_k^- as described in Proposition 2.1, namely,

$$Z_k A_k^- = 0, \qquad A_k Z_k^- = 0. \tag{123}$$

Let $\theta_k : V \subset \mathbb{R}^{n-m} \to U_k \cap C$ denote the corresponding local parametrization of C around x^k defined in (11), such that

$$\theta_k(z_k(x)) = x, \qquad \text{for all } x \text{ in } C; \tag{124}$$

notice that

$$\theta_k(0) = x^k.$$

Theorem 2.2 shows that θ_k is defined on the neighborhood

$$V = B(0, 1/(2\beta\gamma\xi))$$

of the origin in $\mathbb{R}^{n-m}$. In the local coordinate system (122), the tangent space T_k to C at x^k can be represented by $\mathcal{R}(Z_k^-)$,

$$T_k = \{Z_k^- q \mid q \in \mathbb{R}^{n-m}\}, \tag{125}$$

and a natural choice for the Riemannian metric at x^k is

$$\gamma_k(Z_k^- q, Z_k^- q') = \langle q, q' \rangle_{n-m}, \qquad \text{for all } q, q' \in \mathbb{R}^{n-m}. \tag{126}$$

As noticed in Section 3.2, the coordinate system (122) is a normal coordinate system with respect to the Riemannian metric (126), and the geodesic curve $x(\cdot)$ starting from x^k and tangent to

$$v^k = Z_k^- q^k \in T_k$$

is simply

$$x(t) = \theta_k(tq^k); \tag{127}$$

notice that, if

$$\|q^k\|_{n-m} = \alpha_k,$$

the function $x(t)$ is only defined for $t \in [-\tilde{t}, \tilde{t}]$, with

$$\tilde{t} = 1/(2\alpha_k \beta \gamma \xi). \tag{128}$$

We can describe now the various descent methods of Section 4 in the coordinate system (122). According to (66), the gradient of f on C at x^k can be expressed in terms of the *reduced gradient*

$$g^k = Z_k^{-T} \nabla f(x^k). \tag{129}$$

The descent directions p^k defined by (75), (98), (108) can be represented by the general formula

$$p^k = -Z_k^- H_k g^k, \tag{130}$$

where H_k is a symmetric matrix of dimension $n-m$; the direction of steepest descent corresponds to $H_k = I$, the Newton's direction to

$$H_k = Z_k (F_k)^{-1} Z_k^T,$$

and the quasi-Newton direction to

$$H_k = Z_k (G_k)^{-1} Z_k^T.$$

According to (127), the typical iteration is

$$x^{k+1} = \theta_k(\tilde{t}_k(-H_k g^k)), \tag{131}$$

where the stepsize $\tilde{t}_k$ must now be selected on the interval $[0, \tilde{t}]$. Since

$$\alpha_k = \|H_k g^k\|$$

goes to 0, the upper bound $\tilde{t}$, given by (128), increases; and, after a finite number of iterations, $\tilde{t}_k$ coincides with the stepsize t_k defined by any of the selection rules (79), (84), (85), or (86).

Remark 5.1. We can introduce a preconditioning constant matrix D of dimension $n-m$ and define the Riemannian metric by

$$\gamma_k(Z_k q, Z_k^- q') = \langle q, \tfrac{1}{2}(D + D^T) q' \rangle_{n-m}.$$

Remark 5.2. In the Riemannian metric (126), the parallel translation $\pi_0^{\tilde{t}_k} v^k$ of the vector

$$v^k = Z_k^- q^k \in T_k$$

is the vector

$$\tilde{v}^k = Z_{k+1}^- q^k \in T_{k+1};$$

the quasi-Newton equation (113) thus becomes, in the coordinate system

(122),

$$y^k = g^{k+1} - g^k = (H_{k+1})^{-1} s^k, \tag{132}$$

with

$$s^k = -\tilde{t}_k H_k g^k;$$

and the generalized BFGS update formula (116) induces the rank-two update formula for the $(n-m)\times(n-m)$ symmetric matrix H_k^{-1},

$$H_{k+1}^{-1} = H_k^{-1} + \frac{y^k (y^k)^T}{\langle y^k, s^k \rangle} - \frac{H_k^{-1} s^k (s^k)^T H_k^{-1}}{\langle s^k, H_k^{-1} s^k \rangle}. \tag{133}$$

Notice that the sequence of matrices H_k satisfies the formula

$$H_{k+1} = \left[I - \frac{s^k (y^k)^T}{\langle y^k, s^k \rangle} \right] H_k \left[I - \frac{y^k (s^k)^T}{\langle y^k, s^k \rangle} \right] + \frac{s^k (s^k)^T}{\langle y^k, s^k \rangle}. \tag{134}$$

5.2. Tangent-Restoration Approach. The constructive proof of Theorem 2.2 provides us with a practical scheme to compute x^{k+1} as

$$x^{k+1} = \theta_k(\tilde{t}_k p^k) = x^k - \tilde{t}_k Z_k^- H_k g^k + A_k^- w^k, \tag{135}$$

where $w^k \in \mathbb{R}^m$ is chosen such that $x^{k+1} \in C$. Formula (135) can be interpreted geometrically in the following way: x^{k+1} is obtained by a *tangent step* from x^k to

$$\tilde{x}^k = x^k - \tilde{t}_k Z_k^- H_k g^k$$

in the *affine tangent space* $x^k + T_k$, followed by a *restoration step* $A_k w^k$ to enforce feasibility of x^{k+1}. Following (30), the restoration step is determined by taking w^k as the limit of the sequence $w^{k,i}$ starting from

$$w^{k,0} = 0,$$

defined iteratively by

$$w^{k,i+1} = w^{k,i} - c(\tilde{x}^k + A_k^- w^{k,i}), \qquad i = 0, 1, \ldots, \tag{136}$$

since

$$\tilde{t}_k \| p^k \| \leq 1/(2\beta\gamma\xi),$$

the sequence $\{w^{k,i}\}$ has a limit w^k.

The (approximate) geodesic search for the stepsize $\tilde{t}_k$ can then be performed using a finite number of values for the parameter t and computing the restoration step by (136) from the finite number of points

$$\tilde{x}(t) = x^k - t Z_k^- H_k g^k \tag{137}$$

in the affine tangent space. The values for the parameter t can be obtained, starting from an initial guess in $[0, \tilde{t}]$, by an interpolation scheme [e.g., golden section, see Polak (Ref. 17)] to approximate the exact line search (79) or by successive halving to satisfy Armijo's rule (86).

For the steepest descent method ($H_k = I$), it is shown in Gabay and Luenberger (Ref. 1) that an efficient initial guess can be obtained by the first local minimum on $[0, \tilde{t}]$ of the Lagrangian function $l(\tilde{x}(t), \mu^k)$ defined in (61), where

$$\mu^k = -A_k^{-T} \nabla f(x^k)$$

is the approximate Lagrange multiplier (72); the resulting approximation of the idealized steepest descent method converges linearly with the optimal rate given by (88) in Theorem 4.4. A still simpler procedure is presented in Gabay (Ref. 24) and yields the same rate of convergence property.

For the quasi-Newton method [H_k given by update formula (133)], an obvious initial guess is $\min(1, \tilde{t})$. By Theorem 4.7, after a finite number of iterations, the stepsize

$$\tilde{t}_k = 1$$

satisfies Armijo's rule (86), and the method converges superlinearly.

5.3. Efficient Quasi-Newton Methods for Constrained Minimization. We finally present the practical implementations of the quasi-Newton method in the tangent-restoration approach for the specific coordinate systems envisaged in Examples 2.1 and 2.2.

Algorithm 5.1. *Reduced Quasi-Newton Method.* Given $\varepsilon_1, \varepsilon_2 > 0$ (tolerance parameters), $\tau > 0$ (estimate of $\frac{1}{2}\beta\gamma\xi$), $\alpha \in (0, \frac{1}{2})$, $x^k \in C$, H_k^{-1} a definite-positive symmetric matrix of dimension $(n-m)$.

Step 1. Partition the Jacobian matrix $A_k = [B, D]$; compute B^{-1}.

Step 2. Compute the reduced gradient

$$g^k = \nabla_J f(x^k) - D^T (B^{-1})^T \nabla_I f(x^k).$$

Step 3. If $\|g^k\| \leq \varepsilon_1$, then stop; else,

$$\tilde{t} = \|H_k g^k\| \tau;$$

let

$$t = \min(1, \tilde{t}).$$

Step 4. Tangent step:

$$\tilde{x}(t) = x^k - t\begin{bmatrix} -B^{-1}D \\ I \end{bmatrix} H_k g^k, \qquad w^{k,0} = 0.$$

Step 5. Restoration step: while

$$\left\| c\left(\tilde{x}(t) + \begin{bmatrix} B^{-1} \\ 0 \end{bmatrix} w^{k,i}\right)\right\| > \varepsilon_2,$$

do (136),

$$x(t) = \tilde{x}(t) + \begin{bmatrix} B^{-1} \\ 0 \end{bmatrix} w^{k,i}.$$

Step 6. If

$$f(x(t)) \leq f(x^k) - \alpha t \langle g^k, H_k g^k \rangle,$$

then $x^{k+1} = x(t)$, update H_k^{-1} according to (133); else, assign $t/2$ to t, and go to Step 4.

We prefer to update H_k^{-1} according to (133), rather than H_k directly according to (134), because the first scheme is numerically more stable. To compute the tangent direction $H_k g^k$, we then must solve a linear system; since the matrix of this system differs only by a rank-two correction from the one used in the previous iteration, the system can be solved in the order of $(n-m)^2$ multiplications [see Gill and Murray (Ref. 25)].

A similar reduced quasi-Newton method was presented in Gabay and Luenberger (Ref. 1) using a generalized Davidon–Fletcher–Powell update formula; it was, however, acknowledged that its performance was theoretically impaired by the approximate character of the line search in Step 6. The generalized BFGS update offers the superiority of not requiring exact line searches (Theorem 4.7), which makes its choice particularly relevant for Algorithm 5.1.

We now turn to the coordinate system defined in Example 2.2, which is a normal coordinate system with respect to the Riemannian metric on C induced by the Euclidean structure of $\mathbb{R}^n$. It allows us to define a computationally efficient quasi-Newton version of the gradient projection method.

Algorithm 5.2. *Projected Quasi-Newton Method.* Given $\varepsilon_1, \varepsilon_2, \tau, \alpha$, $x^k \in C$, H_k^{-1} defined as in Algorithm 5.1.

Step 1. Factorize the Jacobian matrix

$$A_x = [L, 0]Q;$$

partition

$$Q = \begin{bmatrix} Q_1 \\ Q_2 \end{bmatrix}.$$

Step 2. Compute the Euclidean reduced gradient

$$g^k = Q_2 \nabla f(x^k).$$

Step 3. If

$$\|g^k\| \leq \varepsilon,$$

then stop; else,

$$\tilde{t} = \|H_k g^k\| \cdot \tau;$$

let

$$t = \min(1, \tilde{t}).$$

Step 4. Tangent step:

$$\tilde{x}(t) = x^k - t Q_2^T H_k g^k, \qquad w^{k,0} = 0.$$

Step 5. Restoration step: while

$$\|c(\tilde{x}(t) + Q_1^T L^{-1} w^{k,i})\| > \varepsilon_2,$$

do (136),

$$x(t) = \tilde{x}(t) + Q_1^T L^{-1} w^{k,i}.$$

Step 6. If

$$f(x(t)) \leq f(x^k) - \alpha t \langle g^k, H_k g^k \rangle,$$

then $x^{k+1} = x(t)$, update H_k^{-1} according to (133); else, assign $t/2$ to t, and go to Step 4.

This method generalizes to nonlinear equality constrained problems the Gill and Murray (Ref. 26) version of Goldfarb's method for linearly constrained problems (Ref. 27). Notice that it requires again the updating of the matrix H_k^{-1} of dimension $n-m$ only. The difference between Algorithms 5.1 and 5.2 lies mainly in their steps 1; as noticed in Section 2, the inversion by Gaussian elimination of the basic matrix B requires of the order of $m^3/3$ multiplications, while the QR decomposition of A_k can be obtained with approximately $(n-m/3)m^2$ multiplications. However, the conditioning of the matrices H_k generated by Algorithm 5.2 is generally better than that of the matrices generated by Algorithm 5.1 (which may be severely affected by the chosen partition of A_k).

Remark 5.3. Algorithms 5.1 and 5.2 involve three scalar parameters $\tau, \varepsilon_1, \varepsilon_2$. The estimate τ is introduced to guarantee that the restoration Steps 5 provide a feasible point $x(t)$. In practice, an exact value of τ is unknown; we generally start with a guessed approximation, which is altered to a smaller value if the restoration phase fails in the process of iterations.

The idealized descent methods along geodesics correspond to the choice

$$\varepsilon_1 = \varepsilon_2 = 0.$$

They involve a (generally infinite) sequence of iterations requiring in their restoration step an infinite number of inner iterations. The positive tolerance parameters $\varepsilon_1, \varepsilon_2$ are thus introduced to obtain implementable algorithms which, hopefully, terminate in a finite number of iterations and provide an approximate solution. We must observe that the iterates x^k are then approximately feasible only, i.e.,

$$\|c(x^k)\| < \varepsilon_2,$$

a situation which cannot be handled directly by our convergence theorems. Mukai and Polak (Ref. 28) have given a framework to establish convergence using such approximation schemes; they require essentially that the parameters $\varepsilon_1, \varepsilon_2$ be adapted carefully in the process of iterations to eventually reach an approximate solution.

Remark 5.4. *Simple Arc Search.* In Ref. 1, we proposed, for the implementation of the reduced Newton's Method, to perform the approximation of the arc of geodesic

$$x(t) = \exp_{x^k}(tp^k), \tag{138}$$

using the second-order information available; the sequential tangent-restoration approach is then replaced by an approximate search along a parabola tangent to $Z_k^- p^k$ at x^k, followed by a restoration.

A similar device can be proposed in the framework of the quasi-Newton method along geodesics, involving only first-order information. Suppose, for simplicity, that x^k is close enough to a solution, so that $t = 1$ is a reasonable initial guess for the approximate search along the geodesic given by (138). Let

$$p^k = -H_k g^k,$$

and $q^k \in R^m$, the feasibility correction, be given by

$$q^k = -c(x^k + Z_k^- p^k); \tag{139}$$

and consider the simple arc of curve

$$\tilde{x}(t) = x^k + tZ_k^- p^k + (t^2/2)A_k^- q^k, \qquad \text{for } t \in [0, 1]. \tag{140}$$

Formula (140) defines an arc of parabola starting from x^k and tangent to $Z_k^- p^k$. The points $\tilde{x}(t)$ for $t > 0$ do not in general belong to C, but can be viewed as an approximation of the geodesic obtained by performing only one step of the restoration phase (Step 5).

We thus search the parabola (140) for the first

$$t = 2^{-l}, \qquad l = 0, 1, \ldots,$$

satisfying

$$f(\tilde{x}(t)) \leq f(x^k) + \alpha t \langle g^k, p^k \rangle. \tag{141}$$

A restoration is then performed as in Step 5, and the new iterate defined as in Step 6. This approach presents the advantage of requiring less restoration phases within an iteration.

Remark 5.5. In two recent papers (Refs. 29, 30), Tanabe has proposed and analyzed a continuous version of respectively the projected and the reduced quasi-Newton method. However, he requires the updating of a full $n \times n$ nonsymmetric matrix, which is performed by Broyden's rank-one formula.

6. Problems with Inequality Constraints

We consider now nonlinear programs, where nonlinear inequality constraints appear. For simplicity, we assume that no equality constraints are present. The problem is then given as

$$\min\{f(x) \mid x \in \mathbb{R}^n \text{ s.t. } c_i(x) \leq 0, i = 1, \ldots, p\}, \tag{142}$$

where f and c_i are $\mathscr{C}^\sigma$ real-valued functions on $\mathbb{R}^n$. The constrained set

$$C = \{x \in \mathbb{R}^n \mid c_i(x) \leq 0, i = 1, \ldots, p\} \tag{143}$$

can be embedded in a submanifold of $\mathbb{R}^{n+p}$ in the following way. Define, for $i = 1, \ldots, p$, the $\mathscr{C}^\sigma$ functions $g_i : \mathbb{R}^{n+p} \to \mathbb{R}$ by

$$g_i(x, z) = c_i(x) + z_i^2, \qquad \text{for all } x \in \mathbb{R}^n, z \in \mathbb{R}^p, \tag{144}$$

where z_i is the ith component of z; and consider the set

$$G = g^{-1}(0) \subset \mathbb{R}^{n+p}.$$

Given $x \in C$, the index set of active constraints at x is the subset of

$P=\{1,\ldots,p\}$ defined by

$$I(x)=\{i\in P\,|\,c_i(x)=0\}. \tag{145}$$

Thus, the corresponding point $(x, z)\in G$ is such that

$$z_i=0, \qquad \text{for } i\in I(x),$$

while

$$z_i\neq 0, \qquad \text{for } i\in P-I(x).$$

Let $m(x)$ denote the cardinality of $I(x)$. The Jacobian matrix of the map g is a $p\times(n+p)$ matrix $J_{x,z}$ partitioned as

$$J_{x,z}=[A_x, \Delta_z], \tag{146}$$

where A_x is the $p\times n$ Jacobian matrix of the map c at x and Δ_z is a $p\times p$ diagonal matrix of diagonal entries $2z_i$. Assume that, for all $x\in C$, the gradients of the $m(x)$ active constraints $\nabla c_i(x)$ for $i\in I(x)$ are linearly independent. Then, the matrix $J_{x,z}$ is of full rank p; and, by Theorem 2.1, the set G is a $\mathscr{C}^\sigma$ *differential submanifold* of $\mathbb{R}^{n+p}$ of dimension n.

Problem (142) is thus equivalent to

$$\min_{(x,z)\in G} f(x), \tag{147}$$

minimization of the $\mathscr{C}^\sigma$ differentiable function f over the differential manifold G of class σ. However, notice that the problem is now formulated in the enlarged space $\mathbb{R}^{n+p}$. A naive approach, known in the mathematical programming literature as the *active constraints strategy*, consists in considering at each point $x\in C$ only the active constraints and implementing the previous descent methods in a local coordinate system around x of the (variable) submanifold

$$C(x)=\{x\in\mathbb{R}^n\,|\,c_i(x)=0,\, i\in I(x)\}. \tag{148}$$

It turns out, however, that such a coordinate system cannot be used as a (partial) coordinate system for G around the corresponding point (x, z); in fact, there exist in general points y in a neighborhood of x in $C(x)$ such that

$$c_i(y)>0, \qquad \text{for some } i,$$

i.e., $y\notin C$. In order to apply efficiently the descent methods presented in this paper to problem (142) through formulation (147), we must design convenient local coordinate coordinate systems of G which exploit the separable structure of the functions g_i.

A better approach would consist in adapting directly our methods to the *manifold with boundary* C of R^n by considering, for instance, local coordinate systems mapping a neighborhood $U_x \cap C$ onto a neighborhood of the origin in a finite-dimensional halfspace [see Hirsch (Ref. 14, 1.4)].

7. Conclusions

In this paper, we have developed two distinct themes, a theoretic set-up using the geometry of manifolds and a computation-oriented analysis, and shown how they could be usefully interrelated. We have presented a geometric framework for studying nonlinear programming problems which has enabled us to generalize in an intrinsic manner the analysis and methods of unconstrained minimization to the constrained case. We have then specified the degrees of freedom of this general set-up to our advantage, using the theoretic approach as a guideline for the conception of efficient methods from the computational viewpoint. In particular, we have shown how insights about the structure of the problem could be used to improve the numerical performance of the algorithm by an appropriate choice of the Riemannian structure for the constrained manifold.

We have defined a family of descent methods along geodesics and presented their practical implementation. Such models include most known primal methods for nonlinear programming and some new superlinear converging algorithms.

However, forcing all the iterates to remain on the constraint manifold may be very costly, since it requires one to solve a system of m nonlinear equations at each iteration. This inconvenience is overcome by a new method, formally similar to the reduced quasi-Newton method of this paper, and requiring the feasibility improvement (instead of enforcement) of the successive iterates (Ref. 31).

References

1. GABAY, D., and LUENBERGER, D. G., *Efficiently Converging Minimization Methods Based on the Reduced Gradient*, SIAM Journal on Control and Optimization, Vol. 14, pp. 42–61, 1976.
2. LUENBERGER, D. G., *The Gradient Projection Method along Geodesics*, Management Science, Vol. 18, pp. 620–631, 1972.
3. ROSEN, J. B., *The Gradient Projection Method for Nonlinear Programming: Part II, Nonlinear Constraints*, SIAM Journal on Applied Mathematics, Vol. 9, pp. 514–522, 1961.

4. LICHNEWSKY, A., *Minimisation de Fonctionnelles Définies sur une Variété par la Méthode du Gradient Conjugué*, Thèse de Doctorat d'Etat, Université de Paris-Sud, Paris, France, 1979.
5. ABADIE, J., and GUIGOU, J., *Numerical Experiments with the GRG Method*, Integer and Nonlinear Programming, Edited by J. Abadie, North-Holland Publishing Company, Amsterdam, Holland, pp. 529–536, 1970.
6. MIELE, A., HUANG, H. Y., and HEIDEMAN, J. C., *Sequential Gradient-Restoration Algorithm for the Minimization of Constrained Functions, Ordinary and Conjugate Gradient Versions*, Journal of Optimization Theory and Applications, Vol. 4, pp. 213–243, 1969.
7. MILNOR, J. W., *Topology from the Differential Viewpoint*, University Press of Virginia, Charlottesville, Virginia, 1965.
8. STEWART, G. W., *Introduction to Matrix Computations*, Academic Press, New York, New York, 1973.
9. GUILLEMIN, V., and POLLACK, A., *Differential Topology*, Prentice-Hall, Englewood Cliffs, New Jersey, 1974.
10. MILNOR, J. W., *Morse Theory*, Princeton University Press, Princeton, New Jersey, 1969.
11. HICKS, N. J., *Notes on Differential Geometry*, Van Nostrand Publishing Company, Princeton, New Jersey, 1965.
12. BISHOP, R. L., and CRITTENDEN, R. J., *Geometry of Manifolds*, Academic Press, New York, New York, 1964.
13. LUENBERGER, D. G., *Introduction to Linear and Nonlinear Programming*, Addison-Wesley Publishing Company, Reading, Massachusetts, 1973.
14. HIRSCH, M. W., *Differential Topology*, Springer, New York, New York, 1976.
15. GOLUBITSKY, M., and GUILLEMIN, V., *Stable Mapping and Their Singularities*, Springer, New York, New York, 1973.
16. HESTENES, M. R., *Optimization Theory, The Finite Dimensional Case*, John Wiley and Sons, New York, New York, 1975.
17. POLAK, E., *Computational Methods in Optimization, A Unified Approach*, Academic Press, New York, New York, 1971.
18. ORTEGA, J. M., and RHEINBOLDT, N. C., *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York, New York, 1970.
19. MCCORMICK, G. P., *A Modification of Armijo's Stepsize Rule for Negative Curvature*, Mathematical Programming, Vol. 13, pp. 111–115, 1977.
20. MORÉ, J. J., and SORENSEN, D. C., *On the Use of Directions of Negative Curvature in a Modified Newton Method*, Mathematical Programming, Vol. 16, pp. 1–20, 1979.
21. DENNIS, J. E., and MORÉ, J. J., *Quasi-Newton Methods, Motivation and Theory*, SIAM Review, Vol. 19, pp. 46–89, 1977.
22. POWELL, M. J. D., *Some Global Convergence Properties of a Variable Metric Algorithm for Minimization without Exact Line Searches*, Paper Presented at the AMS/SIAM Symposium on Nonlinear Programming, New York, New York, 1976.

23. DENNIS, J. E., and MORÉ, J. J., *A Characterization of Superlinear Convergence and Its Application to Quasi-Newton Methods*, Mathematics of Computation, Vol. 28, pp. 549–560, 1974.
24. GABAY, D., *Efficient Convergence of Implementable Gradient Algorithms and Stepsize Selection Procedures for Constrained Minimization*, International Computing Symposium 1975, Edited by E. Gelenbe and D. Potier, North-Holland Publishing Company, Amsterdam, Holland, pp. 37–43, 1975.
25. GILL, P. E., and MURRAY, W., *Quasi-Newton Methods for Unconstrained Optimization*, Journal of the Institute of Mathematics and Applications, Vol. 9, pp. 91–108, 1972.
26. GILL, P. E., and MURRAY, W., *Quasi-Newton Methods for Linearly Constrained Optimization*, Numerical Methods for Constrained Optimization, Edited by P. E. Gill and W. Murray, Academic Press, New York, London, 1974.
27. GOLDFARB, D., *Extension of Davidon's Variable Metric Method to Maximization under Linear Inequality and Equality Constraints*, SIAM Journal on Applied Mathematics, Vol. 17, pp. 739–764, 1969.
28. MUKAI, H., and POLAK, E., *On the Use of Approximations in Algorithms for Optimization Problems with Equality and Inequality Constraints*, SIAM Journal on Numerical Analysis, Vol. 13, pp. 674–693, 1978.
29. TANABE, K., *A Geometric Method in Nonlinear Programming*, Journal of Optimization Theory and Applications, Vol. 30, pp. 181–210, 1980.
30. TANABE, K., *Differential Geometry Approach to Extended GRG Methods with Enforced Feasibility in Nonlinear Programming: Global Analysis*, Recent Applications of Generalized Inverses, Edited by M. Z. Nashed, Pitman, London (to appear).
31. GABAY, D., *Reduced Quasi-Newton Methods with Feasibility Improvement for Nonlinearly Constrained Optimization*, Mathematical Programming Studies (to appear).