A quasi-Riemannian approach to constrained optimization

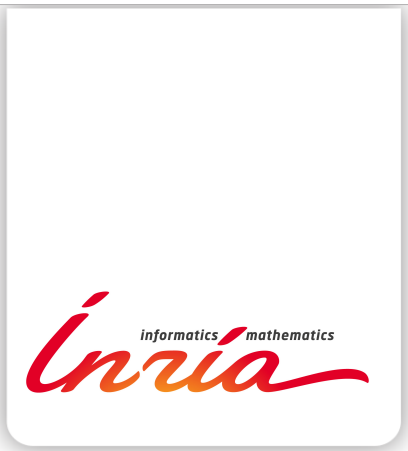
Jean-Antoine Désidéri

A quasi-Riemannian approach to constrained optimization

Jean-Antoine Désidéri

RESEARCH REPORT
N° 9007
December 2016
Project-Team Acumes

ISSN 0249-6399 ISRN INRIA/RR--9007--FR+ENG



A quasi-Riemannian approach to constrained optimization

Jean-Antoine Désidéri*

Project-Team Acumes

Research Report n° 9007 — December 2016 — 16 pages

Abstract: A quasi-Riemannian approach is developed for constrained optimization in which the retraction and transport operators are only approximate. If n is the dimension of the admissible domain, and p the number of scalar equality constraints, the iteration is expressed in terms of a vector of reduced dimension $n-p$ lying in the subspace tangent to the constraint manifold as optimization variable, whereas the minimized function is evaluated at a point, after retraction, that is approximately on the constraint manifold. Precisely, if h is the norm of the tangent vector, the distance between the point of evaluation of the function to be minimized, after retraction, is in general $O(h^4)$, while it would only be $O(h^2)$ if retraction were not applied. The construction only requires evaluation procedures for constraint functions and their gradients to be provided, and eludes the necessity of curvature information.

Key-words: differentiable parametric optimization, nonlinear constraints, Newton's method, BFGS method, Riemannian method

* Directeur de Recherche INRIA, Équipe Acumes

RESEARCH CENTRE
SOPHIA ANTIPOLIS – MÉDITERRANÉE

2004 route des Lucioles - BP 93
06902 Sophia Antipolis Cedex

Une approche quasi-riemannienne en optimisation sous contraintes

Résumé : On développe une approche quasi-riemannienne pour l'optimisation sous contraintes dans laquelle les opérateurs de rétraction et de transport sont seulement approchés. Si n est la dimension de l'espace admissible, et p le nombre de contraintes scalaires d'égalité, l'itération s'exprime en utilisant comme variable d'optimisation un vecteur de dimension $n-p$ du sous-espace tangent aux surfaces de contraintes, alors que la fonction sur laquelle porte l'optimisation est évaluée en un point qui, après rétraction, satisfait approximativement les contraintes. Précisément, si h est la norme du vecteur tangent, la distance entre le point d'évaluation de la fonction à minimiser, après rétraction, est en général $O(h^4)$, alors qu'elle serait seulement $O(h^2)$ sans rétraction. La construction nécessite seulement la donnée de procédures d'évaluation des fonctions de contraintes et de leurs gradients, et aucune information de courbure.

Mots-clés : optimisation paramétrique différentiable, contraintes nonlinéaires, méthode de Newton, méthode BFGS, méthode riemannienne

1 Introduction

In differentiable optimization, the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm [1] is one of the most efficient methods for unconstrained problems. Besides function values, it only requires the specification of the gradient. An approximate Hessian is calculated by successive approximations as part of the iteration, using rank-1 correction matrices. As a result, the iteration has superlinear convergence rate: when minimizing a quadratic function in n variables, if the one-dimensional minimizations in the calculated directions of search are done exactly, the Hessian matrix approximation is exact after n iterations, and from this, the iteration identifies to Newton's iteration, and produces the exact local optimum in only one additional iteration ($n+1$ in total). However the BFGS method does not extend to constrained problems very simply. Following Gabay [2] and other authors, Chunhong Qi et al [3] have proposed a "Riemannian" variant, RBFGS that indeed incorporates equality constraints in the formulation and actually demonstrates superior convergence rates for problems with a large number of variables. However these Riemannian formulations are complicated to implement since they require procedures implementing non-trivial differential-geometry operators ("retraction" and "metric transport") to be developed. In their paper, they assume a formal expression of the constraint to be known. But, in PDE-constrained optimization, many constraints are functional, and it is not clear how can the metric transport operator in particular can be defined.

In this report, a quasi-Riemannian method is defined based on the sole provision of evaluation procedures for the constraint functions and their gradients. By condensing all the equality constraints in one, a purely-explicit approximate retraction operator is defined that yields a point whose distance to the constraint manifold is fourth-order at least. The associated transport operator is approximated. The evaluation of its formal order of accuracy will require further investigation.

2 Method construction

One considers the following constrained-optimization problem:

$$\begin{cases} \min\limits_{\mathbf{x}\in\Omega_a} f(\mathbf{x}) \\ \text{subject to: } g(\mathbf{x}) = 0 \end{cases} \tag{1}$$

where $\Omega_a \subseteq \mathbb{R}^n$ is an open admissible domain in which the function f to be minimized is smooth, say C^2, as well as the p components of function $g:\ \Omega_a \to \mathbb{R}^p$.

The constraint functions are assembled in one single "compound constraint function":

$$\gamma(\mathbf{x}) = \tfrac{1}{2}\mathbf{g}(\mathbf{x})^t g(\mathbf{x}) = \tfrac{1}{2}\sum_{j=1}^{p} g_j(\mathbf{x})^2 \tag{2}$$

The following notations are introduced:

$\mathbf{x}$: starting point; this point is the kth iterate of an outer-loop iteration for the minimization of function f (not specified in this report) by some gradient-based optimization method, such as the BFGS method [1]; by assumption, at this point, the constraints are satisfied: $g_j(\mathbf{x}) = 0$ $(j = 1, \ldots, p)$;

$\mathcal{N}$: subspace orthogonal to the constraint manifold at $\mathbf{x}$; this subspace is spanned by the constraint gradients at $\mathbf{x}$, $\{\nabla g_j(\mathbf{x})\}$ $(j = 1, \ldots, p)$; these vectors are assumed to be linearly independent ("constraint qualification"); through a Gram-Schmidt orthogonalization process, two-by-two orthogonal unit vectors, $\{\omega_{n-p+1}, \ldots, \omega_n\}$ spanning the same subspace are

calculated and the following $n \times p$ matrix is formed:

$$\mathbf{\Omega}_{\mathcal{N}} = \begin{pmatrix} \omega_{1,n-p+1} & \dots & \omega_{1,n} \\ \omega_{2,n-p+1} & \dots & \omega_{2,n} \\ \vdots & \dots & \vdots \\ \omega_{n,n-p+1} & \dots & \omega_{n,n} \end{pmatrix} \tag{3}$$

$\mathcal{T}$: subspace tangent to the constraint manifold at $\mathbf{x}$; $\mathcal{T}$ and $\mathcal{N}$ are supplementary subspaces; $\mathcal{T}$ is defined by completion of the orthonormal basis $\{\omega_j\}$ $(j = 1, \dots, n)$; the tangent subspace is the range of the following $n \times (n-p)$ matrix:

$$\mathbf{\Omega}_{\mathcal{T}} = \begin{pmatrix} \omega_{1,1} & \dots & \omega_{1,n-p} \\ \omega_{2,1} & \dots & \omega_{2,n-p} \\ \vdots & \dots & \vdots \\ \omega_{n,1} & \dots & \omega_{n,n-p} \end{pmatrix} \tag{4}$$

$\bar{\mathbf{x}} = \mathbf{x} + \mathbf{y}$; $\mathbf{y} \in \mathcal{T}$: search variable in the minimization procedure ;

$\bar{\mathbf{x}}^{\perp}$: orthogonal projection of $\bar{\mathbf{x}}$ onto the constraint Riemannian manifold; this theoretical point is not actually calculated in practice;

$\bar{\bar{\mathbf{x}}}_N$: the result of one iteration of the (exact) Newton method for the minimization of function $\gamma(\mathbf{x})$, initiated at $\bar{\mathbf{x}}$ by using the Hessian matrix evaluated at $\bar{\mathbf{x}}$; also not calculated;

$\bar{\bar{\mathbf{x}}}_{QN}$: the result of one iteration of the (approximate) Newton method for the minimization of function $\gamma(\mathbf{x})$ initiated at $\bar{\mathbf{x}}$ by using an approximate Hessian matrix $\bar{\mathcal{H}}$ evaluated at $\bar{\mathbf{x}}$.

These notations are illustrated in Figure 1. This figure has been drawn to scale in a particular case in which f depends on only 2 variables and g is scalar and quadratic.

Note that $\bar{\bar{\mathbf{x}}}_N$ is a theoretical point that will not be computed in practice since the second derivatives of the constraint functions are not assumed to be available. The actual point of evaluation of f will be $\bar{\bar{\mathbf{x}}}_{QN}$. This point will be defined precisely by approximation of these second derivatives as follows.

At $\bar{\mathbf{x}}$, the constraint gradients are again evaluated, as well as $\bar{\mathcal{N}}$ spanned by the local constraint gradients, $\{\nabla g_j(\bar{\mathbf{x}})\}$, and the supplementary subspace $\bar{\mathcal{T}}$ is also identified.

We define the support of vector $\overrightarrow{(\bar{\bar{\mathbf{x}}}_{QN}, \bar{\mathbf{x}})}$ to be $\bar{\mathcal{N}}$, whereas vector $\overrightarrow{(\bar{\mathbf{x}}^{\perp}, \bar{\mathbf{x}})}$ is orthogonal to the constraint manifold at the theoretical point $\bar{\mathbf{x}}^{\perp}$; point $\bar{\bar{\mathbf{x}}}_N$ is situated in the neiborhood of point $\bar{\mathbf{x}}^{\perp}$; however its exact location depends on the unknown first and second derivatives of the constraint functions at $\bar{\mathbf{x}}$, whereas the location of point $\bar{\bar{\mathbf{x}}}_{QN}$ relatively to point $\bar{\mathbf{x}}$ only depends on first derivatives.

Recall the following expressions for the gradient and Hessian of the compound constraint function at some arbitrary point:

$$\nabla\gamma = \sum_j g_j \nabla g_j \;\; \nabla^2\gamma = \sum_j \left[\nabla g_j \left(\nabla g_j\right)^t + g_j \nabla^2 g_j \right]. \tag{5}$$

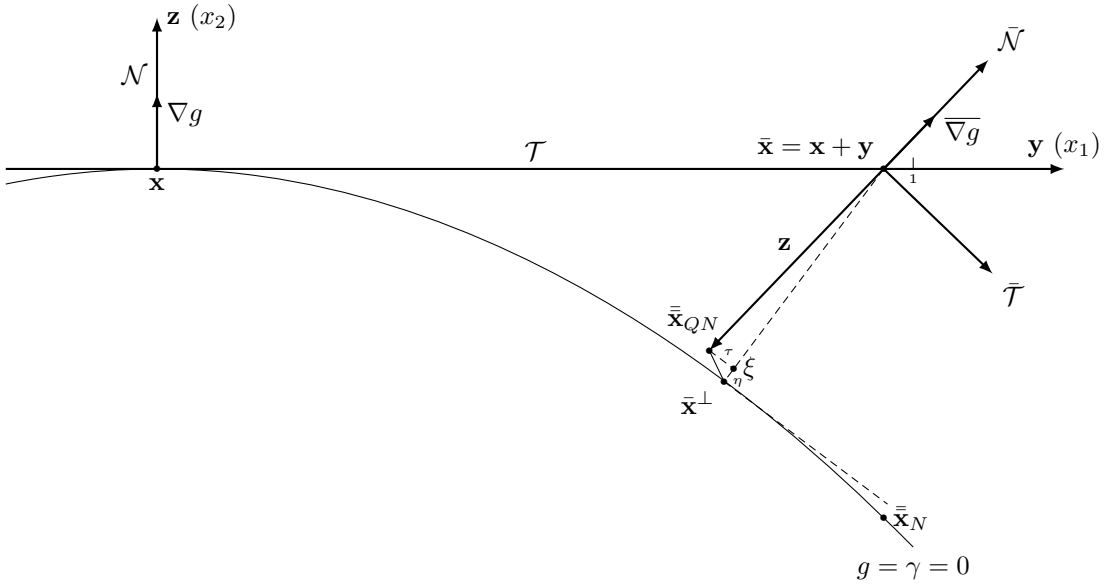
Consequently, specifically at points $\mathbf{x}$ as well as $\bar{\mathbf{x}}^{\perp}$ at which $g = \gamma = 0$, one has:

$$\nabla\gamma = 0 \quad \nabla^2\gamma = \sum_j \left[\nabla g_j \left(\nabla g_j\right)^t \right]. \tag{6}$$

Let:

$$h = \|\bar{\mathbf{x}} - \mathbf{x}\|. \tag{7}$$

(a) Overview



(b) Zoom

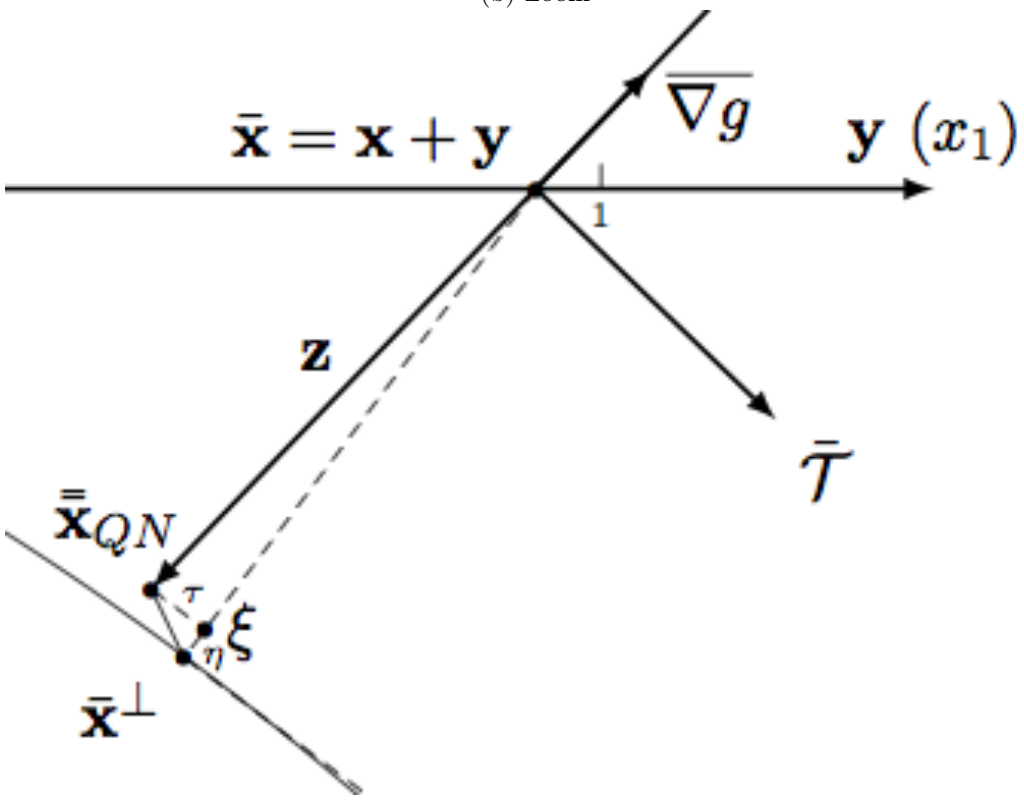


Figure 1: Quasi-Riemannian approach.(This figure has been drawn to scale in the specific case of two optimization variables (x_1, x_2) assuming a single quadratic constraint function $g(x_1, x_2) = \frac{1}{2}x_1^2 + x_2$, a starting point located at $\mathbf{x} = (0, 0)$, and the abscissa of the tangent vector $\mathbf{y}$ set to $y = \bar{x}_1 = \frac{123}{128}$ so that the abscissa of point $\bar{\bar{\mathbf{x}}}_{QN}$ is precisely $\frac{3}{4}$.)

In what follows, an order-of-magnitude analysis is conducted in which h is considered as infinitesimally small.

To elaborate a Riemannian-type optimization method, one should reduce the optimization variable to an $(n-p)$-vector, denoted $\mathbf{y}_{\mathcal{T}}$ hereafter. In this reduction, one retains a vector tangent to the constraint manifold at $\mathbf{x}$:

$$\boxed{\mathbf{y} = \mathbf{\Omega}_{\mathcal{T}}\, \mathbf{y}_{\mathcal{T}}} \tag{8}$$

and **the** $(n-p)$ **components of vector**

$$\boxed{\mathbf{y}_{\mathcal{T}} = (y_1, y_2, \ldots, y_{n-p})^t} \tag{9}$$

are from now on new optimization variables in the course of one (or several) BFGS-type iteration(s) devised to minimize f **in an outer loop using** $\mathbf{x}$ **(at which** $\mathbf{y}_{\mathcal{T}} = 0$**) as starting point.**

One lets:

$$\boxed{\bar{\mathbf{x}} = \mathbf{x} + \mathbf{y}} \tag{10}$$

and one considers the theoretical point $\bar{\bar{\mathbf{x}}}_N$, *not calculated in practice,* that would be obtained by application of one iteration of Newton's method for the minimization of the constraint compound function γ initiated at point $\bar{\mathbf{x}}$, by solving the following equation:

$$0 = \overline{\nabla\gamma} + \overline{\nabla^2\gamma}\,(\bar{\bar{\mathbf{x}}}_N - \bar{\mathbf{x}}) \tag{11}$$

where, and from now on, it is convened to indicate evaluations at $\bar{\mathbf{x}}$ by overlined symbols. Recall that

$$\overline{\nabla^2\gamma} = \sum_j \left(\overline{\nabla g_j}\;\overline{\nabla g_j}^t + \bar{g}_j \overline{\nabla^2 g_j}\right) = \bar{\mathcal{H}} + O(h^2) \tag{12}$$

where one has let:

$$\bar{\mathcal{H}} = \sum_j \overline{\nabla g_j}\;\overline{\nabla g_j}^t. \tag{13}$$

The effective calculation of $\bar{\bar{\mathbf{x}}}_N$ would require the knowledge of the constraint Hessian at $\bar{\mathbf{x}}$, and we precisely exclude this. Instead, we calculate $\bar{\bar{\mathbf{x}}}_{QN}$ from point $\bar{\mathbf{x}}$ by one iteration of a "quasi-Newton's method" defined as follows:

$$0 = \overline{\nabla\gamma} + \bar{\mathcal{H}}\,(\bar{\bar{\mathbf{x}}}_{QN} - \bar{\mathbf{x}})\,. \tag{14}$$

This equation requires some precisions to be made, since matrix $\bar{\mathcal{H}}$ is singular. At point $\bar{\mathbf{x}}$, the constraint gradients $\{\overline{\nabla g_j}\}$ $(j = 1, \ldots, p)$ are assumed to be linearly-independent. They span the subspace $\bar{\mathcal{N}}$. Vector $\overline{\nabla\gamma}$ belongs to this subspace. The range of the linear operator associated with matrix $\bar{\mathcal{H}}$ itself lies in this subspace. **By algorithmic choice,** we force vector

$$\mathbf{z} = \bar{\bar{\mathbf{x}}}_{QN} - \bar{\mathbf{x}} \tag{15}$$

to belong to $\bar{\mathcal{N}}$. So to speak, one solves (14) in this subspace. For this, the orthogonal decomposition of $\mathbb{R}^n = \bar{\mathcal{T}} \oplus \bar{\mathcal{N}}$ is performed at $\bar{\mathbf{x}}$ and the vectors $\{\bar{\omega}_{n-p+1}, \ldots, \bar{\omega}_n\}$ spanning $\bar{\mathcal{N}}$ are defined, and one lets:

$$\mathbf{z} = \sum_{k=1}^{p} \zeta_k \bar{\omega}_{n-p+k} = \mathbf{\Omega}_{\bar{\mathcal{N}}}\, \zeta \tag{16}$$

where:

$$\zeta = (\zeta_1, \zeta_2, \ldots, \zeta_p)^t \tag{17}$$

is the unknown vector. The definition of matrix $\bar{\mathcal{H}}$ given in (13) implies that:

$$\bar{\mathcal{H}}\,(\bar{\bar{\mathbf{x}}}_{QN} - \bar{\mathbf{x}}) = \bar{\mathcal{H}}\mathbf{z} = \sum_{j=1}^{p} \left(\sum_{k=1}^{p} \overline{\nabla g_j}^t \bar{\omega}_{n-p+k} \zeta_k \right) \overline{\nabla g_j} \tag{18}$$

whereas the right-hand side vector in (14) is also expressed in the family of constraint gradients:

$$\overline{\nabla \gamma} = \sum_{j=1}^{p} \bar{g}_j \overline{\nabla g_j}. \tag{19}$$

Hence (14) admits a unique solution in $\bar{\mathcal{N}}$ given by the solution $\mathbf{z}_{\bar{\mathcal{N}}}$ of the following linear system:

$$\forall j: \quad \sum_{k=1}^{p} \overline{\nabla g_j}^t \bar{\omega}_{n-p+k} \zeta_k = -\bar{g}_j \tag{20}$$

or, equivalently:

$$\boxed{\bar{\mathbf{C}}\,\zeta = -\bar{g}} \tag{21}$$

in which matrix $\bar{\mathbf{C}}$ is $p \times p$ and its elements are the following scalar products:

$$\bar{\mathbf{C}}_{j,k} = \overline{\nabla g_j}^t \bar{\omega}_{n-p+k}. \tag{22}$$

This completes the definition of point $\bar{\bar{\mathbf{x}}}_{QN}$, shortly denoted $\bar{\bar{\mathbf{x}}}$ hereafter, whose distance to the constraint manifold will be evaluated in terms of h subsequently. In this way, an approximate "retraction operator" $\bar{\mathbf{x}} \to \bar{\bar{\mathbf{x}}}$ has been defined. To complete the definition of a Riemannian method, here an approximate one, one should also define a "metric transport operator" [2] to evaluate the gradients at $\bar{\bar{\mathbf{x}}}$. A proposition for this is made below, but further investigation is still required to assess it precisely.

Matrix $\bar{\mathbf{C}}$ and vector $\bar{g}$ are evaluated at point $\bar{\mathbf{x}} = \mathbf{x}+\mathbf{y}$. These elements depend on $\mathbf{y} = \mathbf{\Omega}_{\mathcal{T}}\,\mathbf{y}_{\mathcal{T}}$. Hence, by differentiating with respect to $\mathbf{y}$, one obtains:

$$\bar{\mathbf{C}}_{\mathbf{y}}\,\zeta + \bar{\mathbf{C}}\,\zeta_{\mathbf{y}} = -\bar{g}_{\mathbf{y}}. \tag{23}$$

In this report, and prior to further analysis, it is assumed that neglecting the first term in the above, which reflects the variation of the matrix, is legitimate. Accordingly:

$$\bar{\mathbf{C}}\,\zeta_{\mathbf{y}} \doteq -\bar{g}_{\mathbf{y}} \tag{24}$$

that is,

$$\bar{\mathbf{C}}\,d\zeta \doteq -\overline{\nabla g}\,\mathbf{\Omega}_{\mathcal{T}}\,d\mathbf{y}_{\mathcal{T}} \tag{25}$$

and finally:

$$\boxed{\bar{\mathbf{C}}\,\zeta_{\mathbf{y}_{\mathcal{T}}} \doteq -\overline{g'}\,\mathbf{\Omega}_{\mathcal{T}}} \tag{26}$$

in which the Jacobian matrix $\zeta_{\mathbf{y}_{\mathcal{T}}}$ is, by definition, the following:

$$\zeta_{\mathbf{y}_{\mathcal{T}}} = \begin{pmatrix} \zeta_{1,y_1} & \cdots & \zeta_{1,y_{n-p}} \\ \vdots & \ldots & \vdots \\ \zeta_{p,y_1} & \cdots & \zeta_{p,y_{n-p}} \end{pmatrix} \tag{27}$$

and $\zeta_{i,y_j} = \partial \zeta_i / \partial y_j$. Equation (26) provides an approximation of the "transported gradient".

Let us examine the dimensions of the various terms appearing in (21) and (26):

- matrix $\bar{\mathbf{C}}$: $p \times p$;
- vector ζ: $p \times 1$;
- vector $\mathbf{y}_{\mathcal{T}}$: $(n-p) \times 1$;
- Jacobian matrix $\zeta_{\mathbf{y}_{\mathcal{T}}}$: $p \times (n-p)$;
- left-hand-side: $p \times (n-p)$;
- constraint vector g: $p \times 1$;
- position vector $\mathbf{x}$: $n \times 1$;
- Jacobian matrix $\overline{g'} = \overline{\nabla g}$: $p \times n$;
- matrix $\mathbf{\Omega}_{\mathcal{T}}$: $n \times (n-p)$;
- right-hand side: $p \times (n-p)$.

Thus, one solves a linear system defined by the $p \times p$ matrix $\bar{\mathbf{C}}$ and

- one right-hand-side of dimension p in (21), and

- $(n-p)$ right-hand-sides of dimension p in (26),

that is, a total of $n-p+1$ right-hand side vectors.

We now dispose of the elements that are necessary to apply one iteration of the BFGS algorithm (or some other gradient-based algorithm) applied to the restricted function

$$\boxed{\phi(\mathbf{y}_{\mathcal{T}}) := f\big(\bar{\bar{\mathbf{x}}}_{QN}\big) = f\big(\mathbf{x}+\mathbf{y}+\mathbf{z}\big) = f\big(\mathbf{x}+\mathbf{\Omega}_{\mathcal{T}}\,\mathbf{y}_{\mathcal{T}}+\mathbf{\Omega}_{\bar{\mathcal{N}}}\,\zeta\big)} \tag{28}$$

whose gradient is calculated at $\bar{\bar{\mathbf{x}}}_{QN}$ from:

$$\begin{aligned} d\phi(\mathbf{y}_{\mathcal{T}}) &= \nabla f\big(\bar{\bar{\mathbf{x}}}_{QN}\big)^t (d\mathbf{y}+d\mathbf{z}) \\ &= \nabla f\big(\bar{\bar{\mathbf{x}}}_{QN}\big)^t \big[\mathbf{\Omega}_{\mathcal{T}}\,d\mathbf{y}_{\mathcal{T}}+\mathbf{\Omega}_{\bar{\mathcal{N}}}\,d\zeta\big] \\ &= \nabla f\big(\bar{\bar{\mathbf{x}}}_{QN}\big)^t \big[\mathbf{\Omega}_{\mathcal{T}}+\mathbf{\Omega}_{\bar{\mathcal{N}}}\,\zeta_{\mathbf{y}_{\mathcal{T}}}\big]\,d\mathbf{y}_{\mathcal{T}}; \end{aligned} \tag{29}$$

then by transposition:

$$\boxed{\nabla\phi(\mathbf{y}_{\mathcal{T}}) \doteq \big[\mathbf{\Omega}_{\mathcal{T}}^t+\zeta_{\mathbf{y}_{\mathcal{T}}}^t\,\mathbf{\Omega}_{\bar{\mathcal{N}}}^t\big]\nabla f\big(\bar{\bar{\mathbf{x}}}_{QN}\big)} \tag{30}$$

This gradient is of dimension $(n-p)$. One easily checks that the dimensions are compatible. In this expression, the second term reflects the approximation that was made of the constraint manifold curvature.

3 Summary of algorithm

Starting from point $\mathbf{x}$, the following steps are performed:

- Calculation of constraint gradients at $\mathbf{x}$, Gram-Schmidt orthogonalization, orthogonal basis completion, definition of matrices $\mathbf{\Omega}_{\mathcal{T}}$ and $\mathbf{\Omega}_{\mathcal{N}}$, definition of optimisation variables $\mathbf{y}_{\mathcal{T}}$.
- Elaboration of a procedure that performs the following operations for an arbitrary $\mathbf{y}_{\mathcal{T}}$:

- Setting $\bar{\mathbf{x}} = \mathbf{x} + \mathbf{\Omega}_{\mathcal{T}}\, \mathbf{y}_{\mathcal{T}}$
- Calculation of the constraint gradients at $\bar{\mathbf{x}}$, Gram-Schmidt orthogonalization, orthogonal basis completion, definition of matrices $\mathbf{\Omega}_{\bar{\mathcal{T}}}$ and $\mathbf{\Omega}_{\bar{\mathcal{N}}}$.
- Calculation of matrix $\bar{\mathbf{C}}$, and values at $\bar{\mathbf{x}}$ of constraint functions, $\bar{g}$, and gradient $\overline{\nabla g}$. Inversion of the linear system (21)(26) involving a $p \times p$ matrix and $n - p + 1$ right-hand-side vectors.
- Calculation of the restricted function value $\phi(\mathbf{y}_{\mathcal{T}})$ by (29), and its gradient approximation by (30).

Once this procedure is realized, it is possible to apply one or several iterations of the BFGS algorithm applied to the restricted function $\phi(\mathbf{y}_{\mathcal{T}})$, in which the evaluation point remains very close to the constraint manifold as it will now be established.

4 Analysis of orders of magnitude

By Taylor's expansion to first oder of $\nabla\gamma$ about $\bar{\mathbf{x}}$, one obtains:

$$0 = \nabla\gamma(\bar{\mathbf{x}}^{\perp}) = \nabla\gamma(\bar{\mathbf{x}}) + \nabla^2\gamma(\bar{\mathbf{x}})\left(\bar{\mathbf{x}}^{\perp} - \bar{\mathbf{x}}\right) + \underbrace{O(\left\|\bar{\mathbf{x}}^{\perp} - \bar{\mathbf{x}}\right\|^2)}_{O(h^4)} \tag{31}$$

to which (11) is subtracted to get:

$$\nabla^2\gamma(\bar{\mathbf{x}})\left(\bar{\bar{\mathbf{x}}}_N - \bar{\mathbf{x}}^{\perp}\right) = O(h^4). \tag{32}$$

The orders of magnitude of the terms appearing in $\nabla^2\gamma(\bar{\mathbf{x}})$ are the following:

$$\nabla^2\gamma(\bar{\mathbf{x}}) = \sum_j \left[\underbrace{\overline{\nabla g_j}\left(\overline{\nabla g_j}\right)^t}_{O(1)} + \underbrace{\bar{g}_j\overline{\nabla^2 g_j}}_{O(h^2)}\right](\bar{\mathbf{x}}) = \bar{\mathcal{H}} + O(h^2) = O(1) \tag{33}$$

This matrix is singular in $\mathbf{x}$, but invertible in the orthogonal direction $\bar{\mathcal{N}}$. It is admitted that (32) implies that:

$$\boxed{\bar{\bar{\mathbf{x}}}_N - \bar{\mathbf{x}}^{\perp} = O(h^4)} \tag{34}$$

Besides, for (exact) Newton's method (11), (33) implies that:

$$0 = \nabla\gamma(\bar{\mathbf{x}}) + \left[\bar{\mathcal{H}} + O(h^2)\right](\bar{\bar{\mathbf{x}}}_N - \bar{\mathbf{x}}) = \nabla\gamma(\bar{\mathbf{x}}) + \bar{\mathcal{H}}\,(\bar{\bar{\mathbf{x}}}_N - \bar{\mathbf{x}}) + O(h^4) \tag{35}$$

whereas for the quasi-Newton method:

$$0 = \nabla\gamma(\bar{\mathbf{x}}) + \bar{\mathcal{H}}\,(\bar{\bar{\mathbf{x}}}_{QN} - \bar{\mathbf{x}}) \tag{36}$$

Consequently:

$$\bar{\mathcal{H}}\big(\bar{\bar{\mathbf{x}}}_{QN} - \bar{\bar{\mathbf{x}}}_N\big) = O(h^4) \tag{37}$$

and we conclude that:

$$\boxed{\bar{\bar{\mathbf{x}}}_{QN} - \bar{\bar{\mathbf{x}}}_N = O(h^4)} \tag{38}$$

since the inversion of $\bar{\mathcal{H}}$ in the subspace $\bar{\mathcal{N}}$ is well-posed.

By combining (34) and (38), the final conclusion is achieved:

$$\boxed{\boxed{\bar{\bar{\mathbf{x}}}_{QN} = \bar{\mathbf{x}}^{\perp} + O(h^4)}} \tag{39}$$

in which, recall, $\bar{\mathbf{x}}^{\perp}$ is a theoretical point (exactly) on the constraint manifold.

5 Application to a simple case

One considers the simplest case of an optimization in $\mathbb{R}^2$ subject to one scalar constraint given by:

$$g(\mathbf{x}) = ax_1^\alpha + x_2 = 0 \tag{40}$$

where $a > 0$, as in Figure 1 (drawn setting $a = \frac{1}{2}$, $\alpha = 2$). The starting point is $\mathbf{x} = (0, 1)^t$.

The general expression of the gradient is the following:

$$\nabla g = \begin{pmatrix} \alpha a x_1^{\alpha-1} \\ 1 \end{pmatrix}. \tag{41}$$

One considers the following point on the tangent:

$$\bar{\mathbf{x}} = \begin{pmatrix} y \\ 0 \end{pmatrix} \tag{42}$$

where the value of g is the following:

$$\bar{g} = ay^\alpha \tag{43}$$

and the expression of the gradient the following::

$$\overline{\nabla g} = \begin{pmatrix} \alpha a y^{\alpha-1} \\ 1 \end{pmatrix}. \tag{44}$$

Here the subspace $\bar{\mathcal{N}}$ is one-dimensionnal and spanned by the vector

$$\bar{\omega}_2 = \frac{\overline{\nabla g}}{\|\overline{\nabla g}\|} \tag{45}$$

so that

$$\mathbf{z} = \zeta_1 \bar{\omega}_2 \tag{46}$$

and the unique component ζ_1 is the following equation to which (21) here reduces to:

$$\underbrace{\left(\bar{\omega}_2^t \, \overline{\nabla g}\right)}_{\|\overline{\nabla g}\|} \zeta_1 = -\bar{g}. \tag{47}$$

Hence:

$$\mathbf{z} = -\frac{\bar{g}}{\left\|\overline{\nabla g}\right\|^2} \overline{\nabla g} \tag{48}$$

and finally:

$$\bar{\bar{\mathbf{x}}} = \bar{\mathbf{x}} + \mathbf{z} = \begin{pmatrix} \bar{\bar{x}}_1 \\ \bar{\bar{x}}_2 \end{pmatrix} \tag{49}$$

where:

$$\begin{cases} \bar{\bar{x}}_1 = y - \dfrac{ay^\alpha}{1 + \alpha^2 a^2 y^{2\alpha-2}} (\alpha a y^{\alpha-1}) \\ \bar{\bar{x}}_2 = 0 - \dfrac{ay^\alpha}{1 + \alpha^2 a^2 y^{2\alpha-2}} \end{cases} \tag{50}$$

and the value of g at this point is given by:

$$\bar{\bar{g}} = a\bar{\bar{x}}_1^\alpha + \bar{\bar{x}}_2 = a\left[y - \frac{ay^\alpha}{1 + \alpha^2 a^2 y^{2\alpha-2}} (\alpha a y^{\alpha-1})\right]^\alpha - \frac{ay^\alpha}{1 + \alpha^2 a^2 y^{2\alpha-2}} := ay^\alpha \kappa \tag{51}$$

where κ is expressed in terms of the infinitely-small parameter

$$\epsilon = \alpha^2 a^2 y^{2\alpha-2} = O(h^{2\alpha-2}) \tag{52}$$

as follows:

$$\kappa = \left[1 - \frac{\frac{1}{\alpha}\epsilon}{1+\epsilon}\right]^{\alpha} - \frac{1}{1+\epsilon}. \tag{53}$$

Then:

$$\begin{aligned}\kappa &= 1 - \alpha\frac{\epsilon}{\alpha(1+\epsilon)} + \frac{\alpha(\alpha-1)}{2}\left(\frac{\epsilon}{\alpha(1+\epsilon)}\right)^2 + O(\epsilon^3) - \left[1 - \epsilon + \epsilon^2 + O(\epsilon^3)\right] \\ &= \frac{\alpha-1}{2\alpha}\epsilon^2 + O(\epsilon^3) \\ &= O(\epsilon^2)\end{aligned} \tag{54}$$

and finally:

$$\bar{\bar{g}} = O(h^{5\alpha-4}). \tag{55}$$

This indicates a distance of $\bar{\bar{\mathbf{x}}}$ to the constraint manifold in h^6 in the case of a parabolic constraint ($\alpha = 2$) and h^{16} in the case of the quartic constraint ($\alpha = 4$). These results have been confirmed by experiment, see Tables 1-2-3. They are more favorable than the general result given in (39), due to the simplified context of a scalar constraint and two variables.

Case of a parabola ($\alpha = 2$): To simplify one lets $a = \frac{1}{2}$. For values of $y = h_i = 1/2^i$ ($i = 1, \ldots, 4$), the coordinates of $\bar{\bar{\mathbf{x}}}$ are directly calculated from (50), as well as the corresponding value of g.

Table 1: Orders of magnitude estimations in the case of a parabolic constraint

i	h_i	$\bar{\bar{x}}_1$	$\bar{\bar{x}}_2$	$\bar{\bar{g}}_i$	$\bar{\bar{g}}_{i-1}/\bar{\bar{g}}_i$
1	0.5	0.45	-0.1000	0.0013	
2	0.25	0.2426	-0.0294	2.70 10^{-5}	48.1
3	0.125	0.1240	-0.0017	4.66 10^{-7}	58.5
4	0.0625	0.0624	-0.0019	7.39 10^{-9}	63.1

We observe that the ratio of successive $\bar{\bar{g}}$ converges to $64 = 2^6$ as it was established formally.

Case of a quartic constraint ($\alpha = 4$): the experiment was repeated using this time $a = 1$ and $y = h_i = 0.75/1.5^{i-1}$.

Table 2: Orders of magnitude estimations in the case of a quartic constraint

i	h_i	$\bar{\bar{x}}_1$	$\bar{\bar{x}}_2$	$\bar{\bar{g}}_i$	$\bar{\bar{g}}_{i-1}/\bar{\bar{g}}_i$
1	0.75	0.6720	-0.1848	0.0191	
2	0.50	0.4926	-0.0588	0.0001	238
3	0.3333	0.3329	-0.0123	1.38 10^{-7}	583
4	0.2222	0.222195	-0.0024	2.13 10^{-10}	647

This time, we observe that the ratio of successive $\bar{\bar{g}}$ indicates evidence of convergence to $1.5^{16} \doteq 657$, which confirms the formal result of convergence in h^{16}.

We now return to the case of a single scalar parabolic constraint and two variables for a closer examination of the unexpected higher convergence rate in h^6. Another series of experiments of

similar type have been conducted and reported in Table 3. Let us introduce a short notation for the vector

$$\mathbf{v} = \overrightarrow{(\bar{\mathbf{x}}^{\perp}, \bar{\bar{\mathbf{x}}}_{QN})} \tag{56}$$

and denote ξ the projection of point $\bar{\bar{\mathbf{x}}}_{QN}$ onto the direction of $\nabla g(\bar{\mathbf{x}}^{\perp})$, η; let us split vector $\mathbf{v}$ in a component parallel to $\nabla g(\bar{\mathbf{x}}^{\perp})$, η, and the orthogonal component, τ:

$$\mathbf{v} = \tau + \eta \tag{57}$$

($\eta = \overrightarrow{(\bar{\mathbf{x}}^{\perp}, \xi)}$, $\tau = \overrightarrow{(\xi, \bar{\bar{\mathbf{x}}}_{QN})}$, $\eta \perp \tau$). We have established that in general the following bound holds:

$$\|\mathbf{v}\| = O(h^4). \tag{58}$$

Here, we observe that:

$$\|\mathbf{v}\| = O(h^5) \tag{59}$$

which results from

$$\|\tau\| = O(h^5) \quad \|\eta\| = O(h^6). \tag{60}$$

The bound on $\|\eta\|$ itself comes from the fact that the angle between vectors $\mathbf{v}$ and τ is $O(h)$, while the angle between the support of vector $\overrightarrow{(\bar{\bar{\mathbf{x}}}_{QN}, \bar{\mathbf{x}})}$, i.e. $\nabla g(\bar{\mathbf{x}})$, and the support of vector $\overrightarrow{(\bar{\mathbf{x}}^{\perp}, \bar{\mathbf{x}})}$, i.e. $\nabla g(\bar{\mathbf{x}}^{\perp})$, is $O(h^2)$ since the two gradients are evaluated at points that are $O(h^2)$ distant. Consequently, the Taylor expansion to first-oder of $g(\bar{\bar{\mathbf{x}}}_{QN})$ about $\bar{\mathbf{x}}^{\perp}$ writes:

$$g(\bar{\bar{\mathbf{x}}}_{QN}) = \underbrace{g(\bar{\mathbf{x}}^{\perp})}_{0} + \underbrace{\nabla g(\bar{\mathbf{x}}^{\perp}).(\tau+\eta)}_{\substack{\nabla g(\bar{\mathbf{x}}^{\perp}).\eta \text{ car} \\ \tau \perp \nabla g(\bar{\mathbf{x}}^{\perp})}} + \underbrace{O(\|\mathbf{v}\|^2)}_{\substack{\text{au moins } O(h^8), \\ \text{en fait } O(h^{10})}} = O(\|\eta\|) = O(h^6) \tag{61}$$

and produces the observed result. However this is a very particular case.

In order to perform a more general verification, the following procedure was developed to account for situations involving possibly more than two variables and more than one constraint:

- The starting point is $\mathbf{x} = 0$.
- Each constraint is modeled as a quadratic form:

$$g_j(\mathbf{x}) = \tfrac{1}{2}\mathbf{x}^t \mathbf{H}_j \mathbf{x} + b_{n-j+1} x_{n-j+1} \qquad (j = 1, \ldots, p) \tag{62}$$

  where $\{x_i\}$ are the components of vector $\mathbf{x}$. In this way, the subspace tangent to the manifold at $\mathbf{x} = 0$, $\mathcal{T}$, is parameterized as follows $(x_1, \ldots, x_{n-p}, 0, \ldots, 0)$. The $n(n+1)/2$ distinct elements of the Hessian matrix $\mathbf{H}_j$ and the algebraic value of the gradient along its axis, b_{n-j+1}, are drawn using a random generator whose output is transformed to produce a real number of arbitrary sign and absolute value.

- For each random draw defining a complete set of constraint functions, an initial value of h is chosen to be $10^{-5} \times \max_j ||\mathbf{H}_j||_\infty$, and the tangent vector $\mathbf{y}$ (whose norm is h) is then completely defined by angular parameters; these angles are defined in several cases chosen to cover broadly all angular directions.
- Then, for each random draw, and each discretization of the angular parameters completing the definition of $\mathbf{y}$, one computes the point $\bar{\bar{\mathbf{x}}}$, and the corresponding value of the constraint $q_0 = \sqrt{\sum_j \bar{\bar{g}}_j^2}$. Then, the process is repeated for the same set of constraints but after $\mathbf{y}$ has been halved, yielding a compound constraint value q_1; and then $\mathbf{y}$ is halved again, yielding q_2. From $\{q_0, q_1, q_2\}$ the order of convergence is estimated.

For the case of one constraint ($p = 1$) and three variables ($n = 3$), thus two in the tangent plane, in polar coordinates a single angular parameter θ is sufficient. This parameter was given 10 discrete values between 0 and π. For each one, 10 random draws were made of the 7 arbitrary coefficients.

TABLE 3 – Observed orders of magnitude in the parabolic case of two variables $\mathbf{x} = (x_1, x_2)$ and one scalar constraint $g(\mathbf{x}) = \frac{1}{2}x_1^2 + x_2$

i	$h = \bar{\mathbf{x}}_1$	$\bar{\mathbf{x}}_1^{\perp}$	$\bar{\mathbf{x}}_2^{\perp}$	$\bar{\bar{\mathbf{x}}}_{QN_1}$	$\bar{\bar{\mathbf{x}}}_{QN_2}$	$\bar{\bar{g}}_{QN_i}$	$\frac{\bar{\bar{g}}_{QN_{i-1}}}{\bar{\bar{g}}_{QN_i}}$	$\|\mathbf{v}_i\|$	$\frac{\|\mathbf{v}_{i-1}\|}{\|\mathbf{v}_i\|}$	$\|\tau_i\|$	$\frac{\|\tau_{i-1}\|}{\|\tau_i\|}$	$\|\eta_i\|$	$\frac{\|\eta_{i-1}\|}{\|\eta_i\|}$
1	0.9609	0.7500	-0.2812	0.7203	-0.2400	0.2660E-01		0.4569E-01		0.2468E-01		0.3296E-01	
2	0.4805	0.4384	-0.0961	0.4354	-0.0938	0.1015E-02	26.21	0.3734E-02	12.24	0.2926E-02	08.33	0.2107E-02	15.65
3	0.2402	0.2338	-0.0273	0.2337	-0.0273	0.2148E-04	47.26	0.1711E-03	21.82	0.1607E-03	18.45	0.5747E-04	36.65
4	0.1201	0.1193	-0.0071	0.1193	-0.0071	0.3648E-06	58.87	0.6001E-05	28.51	0.5905E-05	27.20	0.1062E-05	54.14
5	0.0601	0.0600	-0.0018	0.0600	-0.0018	0.5824E-08	62.64	0.1933E-06	31.04	0.1926E-06	30.67	0.1734E-07	61.23
6	0.0300	0.0300	-0.0005	0.0300	-0.0005	0.9150E-10	63.66	0.6089E-08	31.75	0.6083E-08	31.66	0.2740E-09	63.29
7	0.0150	0.0150	-0.0001	0.0150	-0.0001	0.1432E-11	63.91	0.1906E-09	31.94	0.1906E-09	31.91	0.4293E-11	63.82
8	0.0075	0.0075	-0.0000	0.0075	-0.0000	0.2238E-13	63.98	0.5961E-11	31.98	0.5960E-11	31.98	0.6712E-13	63.95
9	0.0038	0.0038	-0.0000	0.0038	-0.0000	0.3496E-15	63.99	0.1864E-12	31.98	0.1864E-12	31.98	0.1049E-14	63.96
10	0.0019	0.0019	-0.0000	0.0019	-0.0000	0.5463E-17	64.00	0.5787E-14	32.21	0.5787E-14	32.21	0.1632E-16	64.28
Theoretical extrapolations													
∞	0	0	0	0	0	0	2^6	0	2^5	0	2^5	0	2^6
	h					$O(h^6)$		$O(h^5)$		$O(h^5)$		$O(h^6)$	

The first row of data have been calculated for $h = \frac{123}{128} \doteq 0.9609$, and it cooresponds to the case of Figure 1.
The short notation $\mathbf{v}$ stands for the vector $\mathbf{v} = \bar{\mathbf{x}}^{\perp} - \bar{\bar{\mathbf{x}}}_{QN}$.

Table 3: Observed orders of magnitude in the parabolic case of two variables $\mathbf{x} = (x_1, x_2)$ and one scalar constraint $g(\mathbf{x}) = \frac{1}{2}x_1^2 + x_2$

As an example, below is the output for $\theta = 0$ corresponding to the first random draw:

```
 theta =    0.0000000000000000
 cost, sint =    1.0000000000000000        0.0000000000000000
 idraw =            1
Hmat =
  0.375752 -0.072102  0.424772
 -0.072102 -0.640690  0.354976
  0.424772  0.354976 -0.271549
hnorm =     0.10678E+01
 y =  1.06776740297538923E-005   0.0000000000000000
 b3 =    3.0477759393463062
xbar =    0.00001   0.00000   0.00000
gbar et gradgbar :      0.0000E+00    0.0000   -0.0000    3.0478
i =    1     xbarbar_i =    0.0000
i =    2     xbarbar_i =    0.0000
i =    3     xbarbar_i =   -0.0000
 gbarbar =  -6.70468660565442354E-024
xbar =    0.00001   0.00000   0.00000
gbar et gradgbar :      0.2142E-10    0.0000   -0.0000    3.0478
i =    1     xbarbar_i =    0.0000
i =    2     xbarbar_i =    0.0000
i =    3     xbarbar_i =   -0.0000
 gbarbar =  -4.17629274111245418E-025
xbar =    0.00000   0.00000   0.00000
gbar et gradgbar :      0.5355E-11    0.0000   -0.0000    3.0478
i =    1     xbarbar_i =    0.0000
i =    2     xbarbar_i =    0.0000
i =    3     xbarbar_i =   -0.0000
 gbarbar =  -2.62532909257552729E-026
 Estimated order of convergence by Aitkens quotient =    4.0057577863702223
```

In the above, the notations are straightforwardly understood: theta for θ, cost and sint for $\cos\theta$ and $\sin\theta$, idraw =1 for first draw, Hmat for $\mathbf{H}_1$, hnorm for $||\mathbf{H}_1||_\infty$, y for the 2 components of $\mathbf{y}$, b3 for b_3, xbar for $\bar{\mathbf{x}}$, gbar and gradgbar for $\bar{g}_1$ and $\overline{\nabla g_1}$, xbarbar_i for $\bar{\bar{\mathbf{x}}}_i$, gbarbar for $g_1(\bar{\bar{\mathbf{x}}}) := q$. The second set of data corresponds to halved $\mathbf{y}$, and the third to halved again. The estimated order of convergence is $\alpha = \log_2(q_0 - q_1)/(q_1 - q_2)$.

The results of this experiment are collected in Table 4. They clearly indicate a 4th-order convergence rate. This conclusion confirms the general estimate even in case of a single scalar constraint when the number of variables is at least equal to 3.

θ	1	2	3	4	5	6	7	8	9	10
0	4.0058	4.0000	4.0002	3.9962	3.9988	4.0000	3.9999	4.0003	4.0000	3.9997
$\frac{\pi}{10}$	4.0003	4.0009	4.0001	4.0001	4.0016	4.0003	3.9994	3.9990	3.8585	4.0044
$2\frac{\pi}{10}$	3.9999	3.9997	4.0000	3.9997	3.9955	3.9977	3.9967	4.0001	4.0001	4.0004
$3\frac{\pi}{10}$	4.0004	4.0000	3.9955	4.0020	3.9998	4.0006	3.9999	3.9984	3.9999	4.0010
$4\frac{\pi}{10}$	3.9999	4.0003	3.9985	3.9948	4.0000	3.9982	3.9999	4.0000	3.9996	4.0050
$5\frac{\pi}{10}$	4.0000	4.0001	4.0007	4.0007	3.9995	3.9998	4.0000	4.0009	4.0000	3.9995
$6\frac{\pi}{10}$	4.0000	4.0209	4.0000	4.0000	4.0001	4.0007	4.0001	3.9997	4.0000	4.0008
$7\frac{\pi}{10}$	3.9990	4.0005	4.0002	4.0001	3.9999	4.0000	3.9999	4.0000	4.0000	4.0001
$8\frac{\pi}{10}$	3.9997	4.0001	4.0000	4.0002	4.0000	4.0002	3.9969	4.0005	4.0000	3.9998
$9\frac{\pi}{10}$	3.9999	3.9997	4.0001	4.0001	3.9998	4.0001	4.0004	3.9937	4.0000	3.9971

Table 4: Estimated order of convergence in case of a single scalar quadratic constraint and three variables ($n = 3,\ p = 1$). (For a given θ, the 10 values correspond to different random draws.)

6 Conclusions

The development of a quasi-Riemannian method made in this report will permit to study the application of the BFGS algorithm in case of equality constraints that are not necessarily defined explicitly in terms of the optimization variables, but only through procedures of evaluation of constraint functions and their gradients. This is the case in particular in a setting subject to the solution of a partial-differential equation when the constraints are functional, and evaluating accurate second derivatives is often considered as an unrealistic endeavor. In the present method, no curvature information on the constraint functions is necessary.

We also remark that the singular system, (14), was not found impossible since a particular solution was identified, but under-determined. Thus, it will be examined whether other solutions could yield a higher-order approximation of the constraints.

While the approximation of the retraction operator is considered well established, more analysis of the transport-operator proposed approximation remains necessary.

Acknowledgement

This study was developed in cooperation with ONERA DAAP (*Department of Applied Aerodynamics*) Meudon. In particular, discussions with D. Bailly were extremely fruitful.

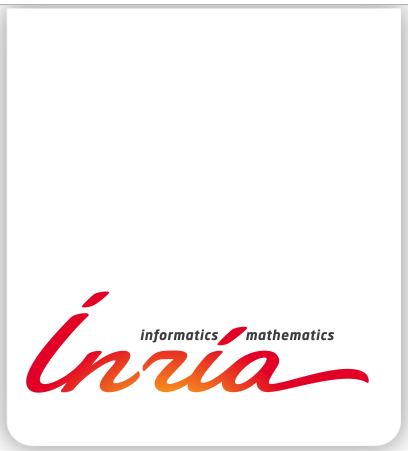
References

[1] Roger Fletcher, *Practical methods of optimization*, 2nd ed., John Wiley & Sons, 1987, `https://en.wikipedia.org/wiki/Broyden-Fletcher-Goldfarb-Shanno_algorithm`.

[2] D. Gabay, *Minimizing a differentiable function over a differential manifold*, J. Optim. Theory Appl. **37** (1982), no. 2.

[3] Chunhong Qi, Kyle A. Gallivan, and P.-A. Absil, *Riemannian BFGS algorithm with applications*, (2010).

Contents



RESEARCH CENTRE
SOPHIA ANTIPOLIS – MÉDITERRANÉE

2004 route des Lucioles - BP 93
06902 Sophia Antipolis Cedex

Publisher
Inria
Domaine de Voluceau - Rocquencourt
BP 105 - 78153 Le Chesnay Cedex
inria.fr

ISSN 0249-6399