# An Efficient BFGS Algorithm for Riemannian Optimization

Chunhong Qi, Kyle A. Gallivan and P.-A. Absil

***Abstract*— In this paper, we present a convergence result for Riemannian line-search methods that ensures superlinear convergence. We also present a theory of building vector transports on submanifolds of $\mathbb{R}^n$ and discuss its use to assess convergence conditions and computational efficiency of the resulting Riemannian optimization algorithms. We illustrate performance and check predictions of our theory using a version of a Riemannian BFGS algorithm we proposed earlier.**

## I. INTRODUCTION

There is a growing interest in the computational mathematics community in optimization problems on manifolds and for efficient algorithms to tackle them that exploit the underlying manifold structure. see, e.g., the recent overview paper [1] and references therein. This paper is part of a research effort to generalize the classical BFGS method to meta-algorithms on abstract Riemannian manifolds backed by detailed convergence analysis, and to show how these meta-algorithms turn into efficient numerical methods for manifold and objective functions of interest.

Some work has been done on BFGS for manifolds. Gabay [2, §4.5] discussed a version using parallel transport. Brace and Manton [3] have a version on the Grassmann manifold for the problem of weighted low-rank approximations. Savas and Lim [4] apply a version on a product of Grassmann manifolds to the problem of best multilinear low-rank approximation of tensors.

In [5] we summarized the five key aspects in which Gabay's Riemannian BFGS [2, §4.5] differs from the classical BFGS method in $\mathbb{R}^n$ (see, e.g., [6]) and presented Riemannian BFGS approach (RBFGS), that is retraction-based and supports the replacement of parallel transport along geodesic with vector transport. The experimental results showed that the version with retraction and vector transport can improve both the convergence and computational cost of Riemannian BFGS. In this paper, we pursue the work started in [5] along the following three directions:

(i) We sketch a crucial step in a forthcoming detailed converge analysis of RBFGS, which involves a Riemannian version of of the Dennis-Moré's condition [7].

(ii) Whereas the notion of retraction has been around for a few years now (see [8]) and is more or less present in several works such as [[9], [10] ], the concept of vector transport is more recent [ [11], §8] and arguably less mature. In this paper, we introduce a new concept, dubbed a *Transporter*, that can be used as a tool to build vector transports on submanifolds of $\mathbb{R}^n$. A given vector transport is not necessarily induced by a transporter, but the two concepts are intimately elated.

(iii) We place the transporter in a projection framework that facilitates choosing an efficient implementation of RBFGS and discussing certain key algebraic properties of the vector transport and its inverse.

The remainder of the paper is organized as follows. Section II briefly reviews the definition of RBFGS from [5] and a convergence theorem is given in Section III. A projection framework is introduced in Section IV and used to define a transporter between tagent spaces of embedded submanifolds. Sufficient conditions are given for a transporter to be a vector transport and to preserve the symmetry of an operator on a tangent space. In section V, examples of the application of the projector framework to implement vector transport and the associated experimental results are shown in Section VI.

## II. A REVIEW OF RBFGS

Given a Riemannian manifold $M$ with Riemannian metric $g$ we assume we have a retraction, a vector transport and its inverse, and a cost function $f$ defined on $M$. The *retraction* on $M$ is a mapping $R$ from the tangent bundle $TM$ onto $M$ and we let $R_x$ denote the restriction of $R$ to $T_xM$ the tangent space of $x$. The notion of retraction on a manifold, due to Adler *et al.* [8], encompasses all first-order approximations to the Riemannian exponential and we refer to [11] or [5] for the specific properties it must satisfy. A vector transport associated with $R$ is a smooth mapping $TM \oplus TM \to TM$, $(\eta_x, \xi_x) \mapsto \mathcal{T}_{\eta_x}\xi_x \in T_{R_x(\eta_x)}M$. The tangent vector $\eta_x \in T_xM$ defines the direction of the transport and, via $R_x$, the tangent space that contains the range of $\mathcal{T}_{\eta_x}$. The vector transport specifies how to move a tangent vector from one tangent space to another. This is also used to move a linear operator from one tangent space to another, e.g., the approximate Hessian. Finally, recall that the gradient of $f$ at $x$, denoted by grad $f(x)$, is defined as the unique element of $T_xM$ that satisfies:

$$g_x(\text{grad } f(x), \xi) = \mathrm{D}f(x)[\xi], \forall \xi \in T_xM.$$

The RBFGS algorithm discussed in [5] is given in Algorithm 1. The algorithm uses the generalization of the Wolfe conditions to $M$ defined by

$$f(R_{x_k}(\alpha_k\eta_k)) \leq f(x_k) + c_1\alpha_k g(\text{grad } f(x_k), \eta_k) \quad (1)$$

$$g\Big((\mathcal{T}_{\alpha_k\eta_k})^{-1}\text{grad } f(R_{x_k}(\alpha_k\eta_k)), \eta_k\Big) \geq c_2 g(\text{grad } f(x_k), \eta_k) \quad (2)$$

Chunhong Qi and Kyle A. Gallivan are with the Department of Mathematics, Florida State University, Tallahassee, Fl 32306, USA {cqi, gallivan}@math.fsu.edu

P.-A.Absil is with the Département d'ingénierie mathématique, Université catholique de Louvain, B-1348 Louvain-la-Neuve, Belgium, absil@inma.ucl.ac.be

ISBN 978-963-311-370-7

with $0 < c_1 < c_2 < 1$. Condition (1) is often called Armijo condition and (2) curvature condition. Other generalizations on $M$ are possible. The generalization of the Euclidean version of BFGS that propagates an approximation to the inverse of the Hessian is also defined in [5].

**Algorithm 1** RBFGS

1: Given: Riemannian manifold $M$ with Riemannian metric $g$; vector transport $\mathcal{T}$ on $M$ with associated retraction $R$; smooth real-valued function $f$ on $M$; initial iterate $\mathbf{x}_0 \in M$; initial Hessian approximation $\mathcal{B}_0$.
2: **for** k = 0, 1, 2, ... **do**
3: Obtain $\eta_k \in T_{\mathbf{x}_k}M$ by solving $\mathcal{B}_k\eta_k = -\text{grad}\, f(\mathbf{x}_k)$.
4: Set step size $\alpha = 1$, $c = g(\text{grad}\, f(\mathbf{x}_k), \eta_k)$. Using line search to find $\alpha_k$ which satisfies conditions (1) and (2). Set $\mathbf{x}_{k+1} = R_{\mathbf{x}_k}(\alpha\eta_k)$.
5: Define $s_k = \mathcal{T}_{\alpha\eta_k}(\alpha\eta_k)$ and

$$y_k = \text{grad}\, f(\mathbf{x}_{k+1}) - \mathcal{T}_{\alpha\eta_k}(\text{grad}\, f(\mathbf{x}_k)).$$

6: Define the linear operator $\mathcal{B}_{k+1} : T_{\mathbf{x}_{k+1}}M \to T_{\mathbf{x}_{k+1}}M$ by

$$\mathcal{B}_{k+1}p = \tilde{\mathcal{B}}_k p - \frac{g(s_k, \tilde{\mathcal{B}}_k p)}{g(s_k, \tilde{\mathcal{B}}_k s_k)}\tilde{\mathcal{B}}_k s_k + \frac{g(y_k, p)}{g(y_k, s_k)}y_k \quad (3)$$

for all $p \in T_{\mathbf{x}_{k+1}}M$, with

$$\tilde{\mathcal{B}}_k = \mathcal{T}_{\alpha\eta_k} \circ \mathcal{B}_k \circ (\mathcal{T}_{\alpha\eta_k})^{-1}. \quad (4)$$

7: **end for**

## III. CONVERGENCE ANALYSIS

A general convergence result for Riemannian optimization algorithms based on Riemannian versions of the idea of line-searches has been presented in [11] in terms of gradient-related sequences of direction vectors and the Armijo condition enforced by Algorithm 1.

*Definition 3.1 ([11]):* Given a cost function $f$ on a Riemannian manifold $M$, a sequence $\{\eta_k\}, \eta_k \in T_{x_k}M$, is gradient related if, for any subsequence $\{x_k\}_{k\in\mathcal{K}}$ of $\{x_k\}$ that converges to a non-critical point of $f$, the corresponding subsequence $\{\eta_k\}_{k\in\mathcal{K}}$ is bounded and satisfies

$$\limsup_{k\to\infty,\, k\in\mathcal{K}} g(\text{grad}\, f(x_k), \eta_k) < 0.$$

The related convergence theorem in [11] can be applied to Algorithm 1 to yield the following result:

*Theorem 3.2 ([11]):* Let $x_0$ be the starting point and $x_k$ be an infinite sequence of iterates generated by Algorithm 1. If the $\eta_k$ in $T_{x_k}M$ are such that the sequence $\{\eta_i\}_{i=0,1,\dots}$ is gradient-relatedand the level set $\mathcal{L} = \{x \in M : f(x) \le f(x_0)\}$ is compact (which holds in particular when $M$ itself is compact) then $\lim_{k\to\infty} \|\text{grad}\, f(x_k)\| = 0$.

So the sequence created by Algorithm 1 converges to a critical point of the cost function if $\eta_k$ is picked such that the sequence $\{\eta_i\}_{i=0,1,\dots}$ is gradient-related and all $\{x_i\}_{i=0,1,\dots}$ are such that the Armijo condition is satisfied with protection against arbitrarily small stepsizes via Armijo backtracking or enforcing both Wolfe conditions. In practice, gradient-related is not a practical requirement to check and constraints are usually imposed on the manner in which the direction vectors are generated to guarantee the condition is satisfied.

While this guarantees convergence, we are interested in achieving acceptably rapid convergence, e.g., superlinear, as is guaranteed with BFGS in $\mathbf{R}^n$. We have generalized an important result from [12, Theorem 8.2.4] that guarantees the basic Riemannian line search algorithm $x_{k+1} = R_{x_k}(\eta_k)$, where $\eta_k = -B_k^{-1}F(x_k)$ converges superlinearly. We impose the requirement on the retraction $R$ that there exist $\mu > 0$, $\tilde{\mu} > 0$ and $\delta > 0$ such that $\forall x \in M$ and $\xi \in T_xM, \|\xi\| \le \delta$

$$\frac{1}{\tilde{\mu}\|\xi\|} \le \text{dist}(x, R_x\xi) \le \frac{1}{\mu\|\xi\|}. \quad (5)$$

*Theorem 3.3:* Let $M$ be a manifold endowed with a $C^2$ vector transport $\mathcal{T}$ and an associated retraction $R$. Let $F$ be a $C^2$ vector field. Also let $M$ be endowed with an affine connection, $\nabla$. Let $\mathbb{D}F(x)$ denote the linear transformation of $T_xM$ defined by $\mathbb{D}F(x)[\xi_x] = \nabla_{\xi_x}F$, where $F$ is a tangent vector field on $M$, $\xi_x$ is a tangent vector to $M$ at $x$. Let $\{B_k\}$ be a sequence of bounded nonsingular linear transformation of $T_{x_k}M$, where $k = 0, 1, \cdots, x_{k+1} = R_{x_k}(\eta_k)$, and $\eta_k = -B_k^{-1}F(x_k)$. Assume that $\mathbb{D}F(x^*)$ is nonsingular, $x_k \ne x^*, \forall k$, and $\lim_{k\to\infty} x_k = x^*$. Then $\{x_k\}$ converges superlinearly to $x^*$ and $F(x^*) = 0$ if and only if

$$\lim_{k\to\infty} \frac{\|[B_k - \mathcal{T}_{\xi_k}\mathbb{D}F(x^*)\mathcal{T}_{\xi_k}^{-1}]\eta_k\|}{\|\eta_k\|} = 0 \quad (6)$$

where $\xi_k \in T_{x^*}M$ is defined by $\xi_k = R_{x^*}^{-1}(x_k)$, i.e. $R_{x^*}(\xi_k) = x_k$.

Theorem 3.3 identifies a key requirement on the evolution of the action of $B_k$ in the direction of $\eta_k$ relative to the action of the covariant derivative. Note that this requirement, like the gradient-related condition above, is quite general and only requires the transport be twice continuously differentiable. In order to apply it to proving the superlinear convergence of RBFGS, we must identify sufficient conditions on the vector transport and the RBFGS iteration that guarantee the required action of $B_k$. In $\mathbf{R}^n$ the fact that the BFGS update preserves symmetry and positive definiteness of the approximate Hessian or approximate inverse is used as a sufficient condition [6]. This is also the case for RBFGS however the preservation condition is more complicated. Proofs for these results on a Riemannian manifold will be given in a forthcoming paper. For the remainder of this paper we concentrate on the efficiency of the vector transport and characterizing when it satisfies the preservation condition.

## IV. TRANSPORT AND SYMMETRY

When considering a submanifold of $\mathbf{R}^n$, tangent spaces are identified with subspaces of $\mathbf{R}^n$ and mappings between subspaces are used to transport vectors and operators between tangent spaces. We have developed a unified point of view of these issues that also lends itself to deriving computationally efficient transport pairs. In this section we consider a projection framework on $\mathbf{R}^n$ and derive sufficient conditions for preserving symmetry of operators mapped

to a different subspace, guaranteeing vector transport and providing computational implementation options.

### A. Linear Mappings and Symmetry

Suppose we are given a subspace $\mathcal{S}$ and an inner product $g(x,y)$ for $x,\ y \in \mathcal{S}$. We can then analyze the symmetry of a linear mapping $A \in \mathbb{R}^{n\times n}$ restricted to $\mathcal{S}$. We have the basis-free characterization of symmetry

*Definition 4.1:* $A \in \mathbb{R}^{n\times n}$ is symmetric with respect to the inner product $g$ on $\mathcal{S}$ if

$$g(PAPx,\ y) = g(x,\ PAPy)$$

where $P$ is a projector onto $\mathcal{S}$.

Symmetry restricted to $\mathcal{S}$ can also be characterized in terms of any basis for $\mathcal{S}$. Suppose the columns of $U_d$, denoted $u_i$, are a basis for $\mathcal{S}$ and for any $x,\ y \in \mathcal{S}$ we write $x = U_d\hat{x}$ and $y = U_d\hat{y}$ for unique $\hat{x},\ \hat{y} \in \mathbb{R}^d$. The inner product $g$ can be written in terms of the basis as

$$g(x,y) = g(U_d\hat{x}, U_d\hat{y}) = \hat{x}^T\hat{G}\hat{y}, \quad \text{where } \hat{e}_i^T\hat{G}\hat{e}_j = g(u_i,u_j)$$

and $\hat{e}_i \in \mathbb{R}^d$ are the standard basis vectors of $\mathbb{R}^d$. Note $\hat{G} = \hat{G}^T$ since the inner product must be commutative. We therefore have

*Definition 4.2:* Given a basis and an inner product $g$ for $\mathcal{S}$, the linear operator $A \in \mathbb{R}^{n\times n}$ is symmetric on $\mathcal{S}$ with respect to $g$ if

$$\hat{A}^T\hat{G} = \hat{G}\hat{A} \quad \text{where } A = U_d\hat{A}U_d^\dagger$$

$\hat{A} \in \mathbb{R}^{d\times d}$ and $U_d^\dagger$ is the generalized inverse that maps $v \in \mathcal{S}$ to the unique $\hat{v} \in \mathbb{R}^d$ such that $v = U_d\hat{v}$ and $\hat{G} \in \mathbb{R}^{d\times d}$ defines $g$ in terms of the basis $U_d$.

If we change the basis from $U_d$ to $\tilde{U}_d = U_dM_d$ where $M_d \in \mathbb{R}^{d\times d}$ is nonsingular the inner product and symmetry is invariant but must be expressed in terms of modified matrices.

We are intersested in preserving symmetry when a symmetric $A$ defined on a subspace, $\mathcal{S}_1$, is mapped to another subspace, $\mathcal{S}_2$. We have the following result.

*Theorem 4.3:* Suppose $(\mathcal{S}_1, g_1)$ and $(\mathcal{S}_2, g_2)$ are inner product spaces with dimension $d$ embedded in $\mathbb{R}^n$ using bases given by the columns of $U_1 \in \mathbb{R}^{n\times d}$ and $U_2 \in \mathbb{R}^{n\times d}$ respectively and the inner products $g_1$ and $g_2$ are defined by $\hat{G}_1 \in \mathbb{R}^{d\times d}$ and $\hat{G}_2 \in \mathbb{R}^{d\times d}$ relative to $U_1$ and $U_2$ respectively Let the linear maps $B_1 : \mathcal{S}_1 \to \mathcal{S}_1$ and $T : \mathcal{S}_1 \to \mathcal{S}_2$ be defined as

$$B_1 = U_1\hat{B}_1U_1^\dagger \in \mathbb{R}^{n\times n},\ \ T = U_2\hat{T}U_1^\dagger \in \mathbb{R}^{n\times n},$$
$$T^\dagger = U_1\hat{T}^{-1}U_2^\dagger \in \mathbb{R}^{n\times n},\ \ \hat{T},\ \hat{B}_1 \in \mathbb{R}^{d\times d}$$

where † indicates the generalized inverse of a matrix. If $B_1$ is symmetric on $(\mathcal{S}_1, g_1)$ and $\hat{G}_1 = (\hat{T}^T\hat{G}_2\hat{T})$ or equivalently $T$ is an isometry, i.e., $g_1(x_1,y_1) = g_2(Tx_1, Ty_1)$ for all $x_1, y_1 \in \mathcal{S}_1$, then the linear map

$$B_2 = TB_1T^\dagger = U_2(\hat{T}\hat{B}_1\hat{T}^{-1})U_2^\dagger = U_2\hat{B}_2U_2^\dagger \in \mathbb{R}^{n\times n}$$

is symmetric on $(\mathcal{S}_2, g_2)$.

It is often the case that $\mathcal{S}_1$ and $\mathcal{S}_2$ inherit their inner products from the inner product on $\mathbb{R}^n$. We then have the following corollary.

*Corollary 1:* Using the definitions of Theorem 4.3, let $U_1$ and $U_2$ be any pair of orthonormal bases for $\mathcal{S}_1$ and $\mathcal{S}_2$ respectively and assume additionally that the inner products $g_1$ and $g_2$ are defined via the inner product $< x,y >= x^TGy$ on $\mathbb{R}^n$. $T$ is an isometry if and only if $\hat{T}^T\hat{T} = I_d$. In which case, $B_2$ is symmetric on $(\mathcal{S}_2, g_2)$.

### B. Vector Transport Theory

The mapping pair $(T, T^\dagger)$, whether isometries or not, are not all vector/inverse vector transport pairs. They must satisfy additional constraints. In this section, we present sufficient condtions for a mapping $T = U_2\hat{T}U_1^\dagger$ to be a vector transport and to be an isometric vector transport on an embedded submanifold of $\mathbf{R}^n$. Proofs will be given in a forthcoming paper. Let $\mathrm{Gr}(d,n)$ denote the Grassmann manifold of $d$-dimensional subspaces of $\mathbf{R}^n$ and $\mathrm{O}_n$ the set of $n \times n$ orthogonal matrices.

*Definition 4.4:* A *transporter* is a smooth (partial) function

$$\ell : \mathrm{Gr}(d,n) \times \mathrm{Gr}(d,n) \to L(\mathbb{R}^n, \mathbb{R}^n),$$

where $L(\mathbb{R}^n, \mathbb{R}^n)$ denotes the set of all linear maps from $\mathbb{R}^n$ into itself, with the following conditions:

1) The domain of definition of $\ell$, denoted by $\mathrm{dom}(\ell)$, contains a neighborhood of the diagonal $\Delta_{\mathrm{Gr}(d,n)} = \{(\mathcal{X},\mathcal{X}) : \mathcal{X} \in \mathrm{Gr}(d,n)\}$.
2) $$\ell(\mathcal{X},\mathcal{Y})\mathcal{X} \subseteq \mathcal{Y}. \tag{7}$$
3) $$\ell(\mathcal{X},\mathcal{Y})\mathcal{X}_\perp = \{0\}. \tag{8}$$
4) Consistency:
$$\ell(\mathcal{X},\mathcal{X})|_{\mathcal{X}} = \mathrm{id}_{\mathcal{X}}, \quad \text{for all } \mathcal{X} \in \mathrm{Gr}(d,n). \tag{9}$$

If moreover $\ell(\mathcal{X},\mathcal{Y})|_{\mathcal{X}}$ is an isometry for all $(\mathcal{X},\mathcal{Y}) \in \mathrm{dom}(\ell)$, where the metric is the one induced from the canonical metric in $\mathbb{R}^n$, then we say that $\ell$ is *isometric*. We say that $\ell$ is *isotropic* if

$$\ell(U\mathcal{X}, U\mathcal{Y}) = U\ell(\mathcal{X},\mathcal{Y})U^T$$

for all $U \in \mathrm{O}_n$; in this case, $\ell$ is fully determined by specifying $\ell(\mathrm{col}(I_{n,d}),\mathcal{Y})$ for all $\mathcal{Y} \in \mathrm{Gr}(d,n)$.

We will abuse notation and write $\ell(X,Y)$ for $\ell(\mathrm{col}(X),\mathrm{col}(Y))$. Let $\mathcal{M}$ denote a manifold endowed with a retraction $R$.

*Theorem 4.5:* If $\ell$ is a transporter, then $\mathcal{T}$ defined by

$$\mathcal{T}_{\eta_n}\xi_x = \ell(T_x\mathcal{M}, T_{R(\eta_x)}\mathcal{M})\xi_x \tag{10}$$

is a vector transport.

In view of (7) and (8), and restricting from now on to orthonormal $X$ and $Y$, we can write

$$\ell(X,Y) = YQ_{X,Y}X^T, \ \text{where } Q_{XM,YN} = N^TQ_{X,Y}M \tag{11}$$

to ensure that $\ell$ induces a function on $\mathrm{Gr}(d,n) \times \mathrm{Gr}(d,n)$ through $\ell(\mathrm{col}(X), \mathrm{col}(Y) = \ell(X,Y)$. The smoothness condition imposes that $(X,Y) \mapsto Q_{X,Y}$ is smooth. The consistency condition imposes that $Q_{X,X} = I$. The mapping $\ell$ is isometric if and only if

$$Q_{X,Y} \in \mathrm{O}_d. \tag{12}$$

Finally, we have the following result that relates fundamental properties of the mapping $\ell$ defined in terms of a specific form of the core operator $Q_{X,Y}$.

*Theorem 4.6:* If $Q$ is defined by

$$Q_{X,Y} = W\rho(\Sigma)V^T, \tag{13}$$

where $Y^TX = W\Sigma V^T$ is an SVD and where $\rho$ is such that, for all signed permutation matrix $P$,

$$P\rho(P^T\Sigma P)P^T = \rho(\Sigma) \tag{14}$$

then isotropy holds for $\ell$ defined through (11). Assuming (13) and (11), consistency holds if and only if $\rho(I) = I$, in which case $\ell$ defines a vector transport through (10). Still assuming (13) and (11), isometry holds if and only if $\rho(\Sigma) \in \mathrm{O}_d$.

This theorem characterizes vector transport and isometric vector transport and therefore can be used with the projection framework to analyze and design efficient vector transport/inverse vector transport pairs.

### C. Projection Framework

Using the point of view of general projection allows us to characterize isometric and nonisometric mappings between subspaces of $\mathbf{R}^n$ in both an analytically and computationally useful manner. Consider two subspaces of $\mathbf{R}^n$ with dimension $d$ and associated bases. We assume $\mathbb{K} = \mathcal{R}(K)$, $\mathbb{L} = \mathcal{R}(L)$ , $\mathbb{K}^\perp = \mathcal{R}(K_\perp)$, $\mathbb{L}^\perp = \mathcal{R}(L_\perp)$, and $\mathbb{K} \neq \mathbb{L}$. Projection yields the decomposition of $\mathbb{R}^n$ and the associated split of the identity matrix

$$\mathbb{K} \oplus \mathbb{L}^\perp = \mathbb{R}^n \quad \text{and} \quad I_n = P + P_\perp$$

We also know by definition

$$\forall z \in \mathbb{R}^n, \quad Pz \in \mathbb{K}, \quad z - Pz \in \mathbb{L}^\perp, \quad P = K(L^TK)^{-1}L^T$$
$$\forall z \in \mathbb{R}^n, \quad P_\perp z \in \mathbb{L}^\perp, \quad z - P_\perp z \in \mathbb{K}, \quad P_\perp = L_\perp(K_\perp^TL_\perp)^{-1}K_\perp^T$$

For computational purposes, we can use either of the two forms for $P$ and $P_\perp$ and choose the most efficient given the relative sizes of $n$ and the dimension of the manifold $d$:

$$P = K(L^TK)^{-1}L^T \quad \text{and} \quad P = I - L_\perp(K_\perp^TL_\perp)^{-1}K_\perp^T$$
$$P_\perp = L_\perp(K_\perp^TL_\perp)^{-1}K_\perp^T \quad \text{and} \quad P_\perp = I - K(L^TK)^{-1}L^T$$

Since $\mathcal{M}$, is an embedded manifold with dimension $d$ in $\mathbb{R}^n$ all elements of the manifold and the tangent bundle are encoded as $n$-vectors. We assume that for each $x \in \mathcal{M}$ we have a matrix $Q_x \in \mathbb{R}^{n\times d}$ such that $T_x = \mathcal{R}(Q_x)$ and $Q_x^TQ_x = I_d$ and a matrix $N_x \in \mathbb{R}^{n\times n-d}$ such that $T_x^\perp = \mathcal{R}(N_x)$ and $N_x^TN_x = I_{n-d}$. The canonical Riemannian metric

$$g(t_1,t_2) =< t_1,t_2 >= t_1^Tt_2$$

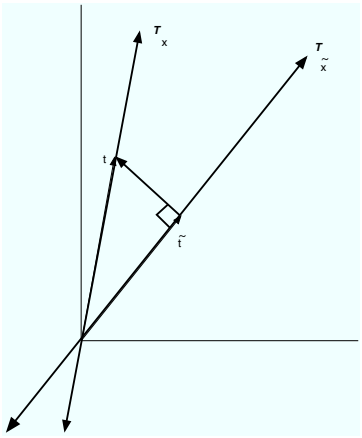


Fig. 1. Orthogonal and oblique projections relating $t$ and $\tilde{t}$

for any $(t_1,t_2) \in T_x \times T_x$ and $x \in \mathcal{M}$ is assumed.

For each $x \in \mathcal{M}$ we need a vector transport, $\mathcal{T} : T_x \to T_{\tilde{x}}$ and inverse vector transport $\mathcal{T}^\dagger : T_{\tilde{x}} \to T_x$ where $\tilde{x} = R_x(\eta_x)$ for some direction vector $\eta_x \in T_x$. These mappings can be represented as $n \times n$ matrices with the forms

$$\mathcal{T} = Q_{\tilde{x}}\hat{T}Q_x^T \quad \text{and} \quad \mathcal{T}^\dagger = Q_x\hat{T}^{-1}Q_{\tilde{x}}^T.$$

Under the assumptions above, taking the core mapping $\hat{T}$ such that $\hat{T}^T\hat{T} = I_d$ guarantees the preservation of symmetry of a transported operator.

The effectiveness of this viewpoint is nicely demonstrated by considering an intuitive choice of mapping that is a vector transport but is not, in fact, an isometry. An orthogonal projection from an arbitrary $v \in \mathbb{R}^n$ to a subspace can be used to define a vector transport.

$$\mathbb{K} = \mathbb{L} = T_{\tilde{x}} = \mathcal{R}(Q_{\tilde{x}}), \quad \mathbb{K}^\perp = \mathbb{L}^\perp = T_{\tilde{x}}^\perp = \mathcal{R}(N_{\tilde{x}})$$
$$P : \mathbb{R}^n \to T_{\tilde{x}}, \quad Pv \mapsto Q_{\tilde{x}}Q_{\tilde{x}}^Tv, \quad P_\perp : \mathbb{R}^n \to T_{\tilde{x}}^\perp, \quad P_\perp v \mapsto N_{\tilde{x}}N_{\tilde{x}}^Tv$$

We can add a projector onto $T_x$ to create the form consistent with our earlier analysis

$$\mathcal{T} = PQ_xQ_x^T = Q_{\tilde{x}}(Q_{\tilde{x}}^TQ_x)Q_x^T = Q_{\tilde{x}}\hat{T}Q_x^T, \quad \mathcal{T}^\dagger = Q_x\hat{T}^{-1}Q_{\tilde{x}}^T.$$

$P$ and $\mathcal{T}$ are equivalent when applied to elements of $T_x$. It is easily verified that $\mathcal{T}$ and $\mathcal{T}^\dagger$ are a vector/inverse vector transport pair on $T_x$ and $T_{\tilde{x}}$. Note, however, that since, in general, $\hat{T}^T\hat{T} \neq I_d$ they are not isometries on $T_x$ and $T_{\tilde{x}}$ and symmetry is not preserved under this choice of transport.

The geometry of the situation with $t \in T_x$, $\tilde{t} = \mathcal{T}t \in T_{\tilde{x}}$, is shown in Figure 1. While $\mathcal{T}$ is an orthogonal projector from $T_x$ to $T_{\tilde{x}}$, $\mathcal{T}^\dagger$ is an oblique projector from $T_{\tilde{x}}$ to $T_x$. Considering Figure 1 yields the intutive notion that taking two oblique projectors using $T_x$, $T_{\tilde{x}}$ and a third space $\mathbb{L}$ common to both projectors and to which both residuals are orthogonal might yield an orthogonal matrix $\hat{T}$. The proposed situation is shown in Figure 2. We have shown that such a space $\mathbb{L}$ always exists under mild assumptions. This yields a pair of isometries that with some care can be made a vector transport/inverse vector transport pair. This is summarized in the following theorem.

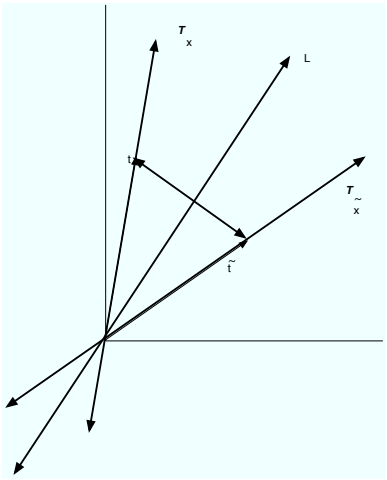


Fig. 2. Two oblique projections relating $t$ and $\tilde{t}$

*Theorem 4.7:* Let $K, \tilde{K} \in \mathbb{R}^{n\times d}$ be such that $K^TK = \tilde{K}^T\tilde{K} = I_d$, $T_x = \mathcal{R}(K)$ and $T_{\tilde{x}} = \mathcal{R}(\tilde{K})$. If $T_x \cap T_{\tilde{x}} = \emptyset$ then for any orthogonal matrix $\hat{T} \in \mathbf{R}^{d\times d}$ there exists $L \in \mathbb{R}^{n\times d}$ with orthonormal columns of the form

$$L = KM + \tilde{K}\tilde{M}$$

with $M = U$, $\tilde{M} = V(Q^TU\Sigma - V)^{-1}(V\Sigma - Q^TU)$ where $K^T\tilde{K} = U\Sigma V^T$. $L$ defines a subspace $\mathbb{L} = \mathcal{R}(L)$ and the associated projectors

$$P = \tilde{K}\hat{T}K^T, \quad \hat{T} = (L^T\tilde{K})^{-1}(L^TK)$$
$$\tilde{P} = K\hat{T}^{-1}\tilde{K}^T, \quad \hat{T}^{-1} = (L^TK)^{-1}(L^T\tilde{K})$$
such that
$$P\tilde{P} = \tilde{K}\tilde{K}^T \quad \text{and} \quad \tilde{P}P = KK^T.$$

The projectors define a transform and its inverse between subspaces $T_x$ and $T_{\tilde{x}}$ that are isometries and given the operators $A : T_x \to T_x$ and $\tilde{A} = PA\tilde{P} : T_{\tilde{x}} \to T_{\tilde{x}}$ the symmetry of $A$ on $T_x$ implies the symmetry of of $\tilde{A}$ on $T_{\tilde{x}}$ and vice versa.

Theorem 4.7 assumes that $T_x \cap T_{\tilde{x}} = \emptyset$. This is not a significant limitation. When there is a nontrivial intersection the components of tangent vectors in the intersection can be left unaltered by the vector transport and its inverse. This saves computation and enforces consistency on the intersection as required by the definition of vector transport.

Two useful isometric vector transports are easily defined. We assume $U_1 \in \mathbb{R}^{n\times d}$ and $U_2 \in \mathbb{R}^{n\times d}$ with $\hat{G}_1 = U_1^TGU_1 = U_2^TGU_2 = \hat{G}_2 = I_d$, and $\mathcal{S}_1 = \mathcal{R}(U_1)$ and $\mathcal{S}_2 = \mathcal{R}(U_2)$ where the inner product is defined by $G$ and inherited on all subspaces. Consider the linear maping $T : \mathbb{R}^n \to \mathcal{S}_2$ and its inverse defined by projection given by

$$T = U_2U_2^\dagger U_1U_1^\dagger = U_2\hat{T}U_1^\dagger \text{ and } T^\dagger = U_1\hat{T}^{-1}U_2^\dagger$$

Canonical bases for $\mathcal{S}_1$ and $\mathcal{S}_2$ can be used to define an isometric vector transport. Let $U_2^TGU_1 = W\Sigma V^T$ and $\hat{T} = WV^T$. We have $W^TU_2^TGU_1V = \Sigma = \tilde{U}_2^TG\tilde{U}_1$ and

$$T = U_2\hat{T}U_1^\dagger = \tilde{U}_2\tilde{U}_1^\dagger \tag{15}$$

$\tilde{U}_1$ and $\tilde{U}_2$ are the canonical bases with respect to the inner product defined by $G$ and $T$ is an isometry.

The economical $QR$ factorization defines an isometric vector transport that is less expensive computationally. If $G = I_n$ defines the inner product then the mapping

$$T = \text{qf}(U_2U_2^TU_1)U_1^T = \tilde{U}_2U_1^T \tag{16}$$

where qf($A$) is the rectangular factor with orthonormal columns in the economical $QR$ factorization of $A$ is a vector transport. This is easily generalized to the case where $G \neq I_n$.

The projection framework also gives us a set of formulations of the pair of mappings from which we may build multiple computational versions. For example, for the nonisometric vector transport above, computationally we do not need to include the $Q_xQ_x^T$ factor added to $P$ to form $\mathcal{T}$ for analytical purposes if the input is restricted to vectors in $T_x$. So we can take the computational form of $\mathcal{T}$ to be $T = Q_{\tilde{x}}Q_{\tilde{x}}^T$ and the projection framework gives us the straightforward computational choices built from the simple identities

$$T = Q_{\tilde{x}}Q_{\tilde{x}}^T \quad \text{and} \quad T = I - \mathcal{T}_\perp = I - N_{\tilde{x}}N_{\tilde{x}}^T$$
$$T_\perp = N_{\tilde{x}}N_{\tilde{x}}^T \quad \text{and} \quad T_\perp = I - Q_{\tilde{x}}Q_{\tilde{x}}^T$$

For isometries Theorem 4.7 characterizes the spaces and the associated forms of the oblique projectors that can be used similarly to enumerate computational possibilities.

## V. IMPLEMENTATION ON THE UNIT SPHERE

We have discussed two forms of implementation of RBFGS in [5] that differ based on the use of bases for tangent spaces and the manner in which $B_k$ and associated mappings are represented, i.e., as $n \times n$ matrices and without using bases or in terms of their $d\times d$ core matrices and bases. These forms assume that efficient computational representations are available for $\mathcal{T}$ and $\mathcal{T}^\dagger$ as matrices and for their application to tangent vectors via $\mathcal{T}t$ and $\mathcal{T}^\dagger\tilde{t}$. In some cases, only the latter applications are available in efficient form in which case two hybrid implementation approaches are possible that selectively use bases and factorizations of $B_k$ and associated matrices. One requires an efficient $\mathcal{T}t$ only while the other requires both $\mathcal{T}t$ and $\mathcal{T}^\dagger\tilde{t}$. In this section we present some examples of these operations.

### A. Nonisometric Vector Transport

We view the unit sphere $S^{n-1} = \{x \in \mathbb{R}^n : x^Tx = 1\}$ as a Riemannian submanifold of the Euclidean space $\mathbb{R}^n$ with the inherited inner product on each tangent space. The tangent space at $x$, orthogonal projection onto the tangent space at $x$, and the retraction chosen are given by

$$T_xS^{n-1} = \{\xi \in \mathbb{R}^n : x^T\xi = 0\}$$
$$\text{P}_x\xi = \xi - xx^T\xi$$
$$R_x(\eta_x) = (x + \eta_x)/\|(x + \eta_x)\|,$$

where $\|\cdot\|$ denotes the Euclidean norm. Denoting $R_x(\eta_x) = \tilde{x}$, we have the following

$$\begin{aligned} T_x &= \mathcal{R}(Q_x), \quad Q_x^T Q_x = I_d \\ T_{\tilde{x}} &= \mathcal{R}(Q_{\tilde{x}}), \quad Q_{\tilde{x}}^T Q_{\tilde{x}} = I_d \\ N_x &= x, \quad \mathcal{N}_x = \mathcal{R}(N_x) \\ N_{\tilde{x}} &= \tilde{x}, \quad \mathcal{N}_{\tilde{x}} = \mathcal{R}(N_{\tilde{x}}) \end{aligned}$$

Applying the projection framework we have the options

$$\mathcal{T} = Q_{\tilde{x}} Q_{\tilde{x}}^T \text{ or } \mathcal{T} = I - \mathcal{T}_\perp = I - \tilde{x}\tilde{x}^T$$

$$\mathcal{T}^\dagger = Q_x (Q_{\tilde{x}}^T Q_x)^{-1} Q_{\tilde{x}}^T \text{ or } \mathcal{T}^\dagger = I - \tilde{x}(x^T\tilde{x})^{-1}x^T$$

So from a complexity point of view we use the latter form of each to define a projection-based nonisometric vector transport pair since they involve only outer products which will lead to an $O(n)$ complexity when applying $\mathcal{T}$ and $\mathcal{T}^\dagger$ to a vector and $O(n^2)$ when applying $\mathcal{T}$ and $\mathcal{T}^\dagger$ to an $n \times n$ matrix.

### B. Isometric Vector Transport

Isometric vector transports based on canonical angles and the economical *QR* factorization can be implemented directly based on (15) and (16) if the appropriate bases are generated. These are very inefficient compared to the form that can be derived by applying Theorem 4.7 and considering the various forms possible.

For the unit sphere $\mathcal{I} = T_x \cap T_{\tilde{x}}$ is a subspace of dimension $n-2$ so it is useful computationally to exploit knowledge of the structure of the embedded manifold to derive an implementation with complexity comparable to that of the nonisometric transport implementation above.

We use the following decompositions of the spaces

$$\begin{aligned} \mathcal{I} = T_x \cap T_{\tilde{x}}, \quad T_x = C_x \oplus \mathcal{I}, \quad T_{\tilde{x}} = C_{\tilde{x}} \oplus \mathcal{I} \\ \mathbb{R}^n = \mathcal{I}^\perp \oplus \mathcal{I}, \quad \mathcal{I}^\perp = \mathcal{N}_x \oplus C_x = \mathcal{N}_{\tilde{x}} \oplus C_{\tilde{x}} \end{aligned}$$

We assume that $0 < x^T\tilde{x} < 1$ and define the bases

$$\begin{aligned} \mathcal{I}^\perp = span[x,\ \tilde{x}] = span[x,\ q] = span[\tilde{q},\ \tilde{x}] \\ \tilde{r} = (I - \tilde{x}\tilde{x}^T)x, \quad \tilde{q} = \tilde{r}/\|\tilde{r}\|_2 \\ r = (I - xx^T)\tilde{x}, \quad q = r/\|r\|_2 \\ C_x = \mathcal{R}(q), \quad C_{\tilde{x}} = \mathcal{R}(\tilde{q}) \end{aligned}$$

It can be shown that the sign of $x^T\tilde{x}$ is opposite to that of $q^T\tilde{q}$ when using the formulas above. In fact, $q^T\tilde{q} = -x^T\tilde{x}$. So in order to guarantee that we have a vector transport, i.e., consistent and continuous, we change the sign of either $q$ or $\tilde{q}$ before proceeding with the calculations below.

For any $t \in T_x$ and $\tilde{t} \in T_{\tilde{x}}$ where $\mathcal{T}t = \tilde{t}$ and $t = \mathcal{T}^\dagger\tilde{t}$ we have

$$\begin{aligned} t = t_c + t_\cap, \quad \tilde{t} = \tilde{t}_c + t_\cap \\ t_c \in C_x, \quad \tilde{t}_c \in C_{\tilde{x}}, \quad t_\cap \in \mathcal{I} \end{aligned}$$

i.e., they share the component in $\mathcal{I}$ and the components in $\mathcal{I} - \mathcal{N}_x$ and $\mathcal{I} - \mathcal{N}_{\tilde{x}}$ are related by vector transport

$$\tilde{t}_c = \hat{\mathcal{T}} t \quad \text{and} \quad t_c = \hat{\mathcal{T}}^\dagger \tilde{t}_c$$

Theorem 4.7 can be applied to determine an efficient form of $\hat{\mathcal{T}}$ and $\hat{\mathcal{T}}^\dagger$ by determining the space $\mathbb{L} = \mathcal{R}(\ell_1)$. It results that the vector transports betwen $C_x$ and $C_{\tilde{x}}$

$$\begin{aligned} \hat{\mathcal{T}} &= \tilde{q}(\ell_1^T\tilde{q})^{-1}\ell_1^T q q^T = \tilde{q}(\ell_1^T\tilde{q})^{-1}(\ell_1^T q)q^T = \tilde{q}q^T \\ \hat{\mathcal{T}}^\dagger &= q(\ell_1^T q)^{-1}\ell_1^T\tilde{q}\tilde{q}^T = q(\ell_1^T q)^{-1}(\ell_1^T\tilde{q})\tilde{q}^T = q\tilde{q}^T \end{aligned}$$

where $\ell_1 = \mu(q + \tilde{q})$ with $\mu$ is a scale to normalize the length of $\ell_1$. The maps are clearly isometries.

For any $t \in T_x$ we have $t = t_c + t_\cap$ and

$$\begin{aligned} \tilde{t} &= t_\cap + \hat{\mathcal{T}} t_c = t_\cap + \tilde{q}q^T t \\ &= t - \tilde{x}\tilde{x}^T t - \tilde{q}\tilde{q}^T t + \tilde{q}q^T t \end{aligned}$$

Similarly for $\tilde{t} \in T_{\tilde{x}}$ we have $\tilde{t} = \tilde{t}_c + t_\cap$ and

$$\begin{aligned} t &= t_\cap + \hat{\mathcal{T}}^\dagger\tilde{t}_c = (q\tilde{q}^T)\tilde{q}\tilde{q}^T\tilde{t} + t_\cap \\ &= \tilde{t} - \tilde{x}\tilde{x}^T\tilde{t} - \tilde{q}\tilde{q}^T\tilde{t} + q\tilde{q}^T\tilde{t} \end{aligned}$$

Note these transports have complexity only slightly higher than the nonisometric forms differing in the constant not the order.

Finally, for the unit sphere, the Levi-Civita parallel transport of $t \in T_x$ along the geodesic, $\gamma$, from $x$ in direction $\eta \in T_x$ is [13] can be written in the efficient form

$$P_\gamma^{t \leftarrow 0}\xi = \Big(I_n + (\cos(\|\eta\|t) - 1)\frac{\eta\eta^T}{\|\eta\|^2} - \sin(\|\eta\|t)\frac{x\eta^T}{\|\eta\|}\Big)\xi.$$

This parallel transport and its inverse, which are of course isometries, have computational costs comparable to the efficient forms of the vector transports and their inverses.

## VI. EXPERIMENT

In this section we present the results of a simple problem on the unit sphere to verify our predictions based on the discussions above. We consider the minimization of the Rayleigh quotient on the unit sphere. For a symmetric matrix $A$, the unit-norm eigenvector, $v$, corresponding to the smallest eigenvalue, defines the two global minima, $\pm v$, of the Rayleigh quotient $f : S^{n-1} \to \mathbb{R}, x \mapsto x^T Ax$. The gradient of $f$ is given by

$$\text{grad } f(x) = 2\mathrm{P}_x(Ax) = 2(Ax - xx^T Ax)$$

Table I shows the number of iterations and time required for RBFGS to reduce the norm of the gradient of the Rayleigh Quotient cost function below $10^{-5}$ using the efficient nonisometric vector transport derived above (NI), the inefficient form of the canonical-bases isometric vector transport (CB) of (15), the equivalent but computationally efficient form of the canonical-bases isometric vector transport (CBE) derived above and the inefficient form of the *QR*-based isometric vector transport of (16) (QR).

As expected, the isometric vector transports converge at the same rate, while the efficient isometric vector transport derived by using the variety of implementations apparent from the projection framework uses less time than the inefficent versions. Note the efficiency and effectiveness of the nonisometric vector transport. This demonstrates that the preservation of symmetry on each step of RBFGS using

TABLE I
PERFORMANCE OF SEVERAL VECTOR TRANSPORTS

| Rayleigh $n = 300$ | | | | |
|---|---|---|---|---|
| | NI | CB | CBE | QR |
| Time (sec.) | 4.0 | 20 | 4.7 | 15.8 |
| Iteration | 97 | 92 | 92 | 97 |

an isometric vector transport is only a sufficient condition and that nonisometries can be competitive and satisfy the necessary and sufficient conditions of Theorem 3.3. This aspect of RBFGS will be considered in a future paper.

## VII. CONCLUSION

In this paper we have summarized a generalization to Riemannian manifolds of a convergence theorem for Euclidean line search methods for optimization. We have also described a projection framework for characterizing vector transports and identifying those that preserve symmetry and positive definiteness of linear mappings transported between tangent spaces of an embedded submanifold of $\mathbf{R}^n$. The computational efficiency implications of the projection framework have also been discussed and illustrated on a simple example on the unit sphere.

## VIII. ACKNOWLEDGMENTS

This paper presents research results of the Belgian Network DYSCO (Dynamical Systems, Control, and Optimization), funded by the Interuniversity Attraction Poles Programme, initiated by the Belgian State, Science Policy Office. The scientific responsibility rests with its authors.

## REFERENCES

[1] P.-A. Absil, R. Mahony and R. Sepulchre, "Optimization on manifolds: methods and applications", *Recent Advances in Optimization and its Applications in Engineering*, Springer, 2009.

[2] D. Gabay, "Minimizing a differentiable function over a differential manifold", *J. Optim. Theory Appl.*, 37(2):177–219, 1982.

[3] Ian Brace and Jonathan H. Manton, "An improved BFGS-on-manifold algorithm for computing weighted low rank approximations", *In proceedings of the 17h international symposium on mathematical theory of networks and systems*, pages 1735–1738, 2006.

[4] Berkant Savas and Lek-Heng Lim, "Best multilinear rank approximation of tensors with quasi-newton methods on grassmannians". *Technical Report LITH-MAT-R-2008-01-SE*, Department of Mathematics, Linköpings University, 2008.

[5] Chunhong Qi, Kyle A. Gallivan, P.-A. Absil, "Riemannian BFGS algorithm with applications", *Recent Advances in Optimization and its Applications in Engineering*, Springer, 2009.

[6] Jorge Nocedal and Stephen J. Wright, "Numerical optimization", *Springer Series in Operations Research and Financial Engineering*, Springer, New York, second edition, 2006.

[7] J.E.Dennis and J.More, "Quasi-newton methods, motivation and theory", *TR*, 74-217, 1974.

[8] Roy L. Adler, Jean-Pierre Dedieu, Joseph Y. Margulies, Marco Martens, and Mike Shub, "Newton's method on Riemannian manifolds and a geometric model for the human spine", *IMA J. Numer. Anal.*, 22(3):359–390, 2002.

[9] Manton, Jonathan H., "Optimization algorithms exploiting unitary constraints", *IEEE Trans. Signal Process.*, volume 50, pp. 635-650, 2002.

[10] E. Celledoni and A. Iserles, "Methods for the approximation of the matrix exponential in a Lie-algebraic setting", *IMA J. Numer. Anal.*, 21(2):463–488, 2001.

[11] P.-A. Absil, R. Mahony, and R. Sepulchre, "Optimization Algorithms on Matrix Manifolds", *Princeton University Press*, Princeton, NJ, 2008.

[12] John E. Dennis, Jr. and Robert B. Schnabel, "Numerical methods for unconstrained optimization and nonlinear equations", *Prentice Hall Series in Computational Mathematics*, Prentice Hall Inc., Englewood Cliffs, NJ, 1983.

[13] N.Del Buono and C.Elia, "Computation of few Lyapunov exponents by geodesic based algorithms", *Future Generation Computer systems*, volume 19, pp. 425-430, 2003.