Yokohama Publishers
ISSN 1349-8169 ONLINE JOURNAL

Pacific Journal of Optimization
Vol. 15, No. 2, pp. 279 – 294, 2019

A SUBSPACE METHOD FOR OPTIMIZATION ON RIEMANNIAN MANIFOLDS*

Hejie Wei

Abstract: In this paper, the subspace method proposed by Yuan and Stoer (ZAMM Z. Angew. Math. Mech. 75 (1995) 69-77) for euclidean unconstrained optimization is generalized for optimization on Riemannian manifolds. Under some conditions, the global convergence and local linear convergence of the algorithm are established. Our numerical results show that the proposed algorithm is competitive with some recent developed methods.

Key words: *subspace method, Riemannian manifold, vector transport*

Mathematics Subject Classification: *90C30, 65K05*

1 Introduction

Recently, there are many research works on the topic of optimization problems on Riemannian manifolds, because of its applications in various areas, such as signal processing , neural networks, computer vision, and econometrics, see, for instance [3, 14, 16, 19, 21]. Traditional optimization algorithms for solving smooth problems, such as steepest descent, Newton, quasi-Newton, nonlinear conjugate gradient and trust-region method have been successfully generalized to Riemannian manifolds. See [1–3, 5–7, 9, 13, 16] and the references therein. For the theory and algorithms for nonsmooth optimization on Riemannian manifolds, see, for instance [10, 11, 17, 18, 22] .

The realistic application problems of optimization on Riemannian manifolds are always large scale problems. Therefore, some techniques for handling large scale problems on Euclidean spaces obtain much attention. See [14, 19, 21] and the references therein. Among these, the subspace techniques are getting more and more important. Absil and Gallivan presented accelerated line-search and trust region methods in [4]. In this paper, we propose a subspace method which is a generalization of the subspace method proposed by Yuan and Stoer [20]. In Yuan and Stoer's method, a search direction is computed by minimizing the approximate quadratic model in the two dimensional subspace spanned by the current gradient and the last search direction. We also adopt the same idea to compute the search direction which lies in a two dimensional subspace. Under some mild conditions, we prove the global convergence and local linear convergence of our method.

The paper is organized as follows. In section 2, we introduce some notations, definitions and preliminary results that will be frequently used in the subsequent discussions. In section

*This research is supported by the National Natural Science Foundation of China NSFC-11371102.

3, we summarize Yuan and Stoer's method briefly and propose our subspace method in detail. In section 4, the global and local linear convergence are established. Numerical experiments are reported in section 5.

2 Notations, Definitions and Preliminary Results

In this section, we introduce the notations and definitions which will be used throughout the paper. For any $x, y \in \mathbb{R}^n$, the inner product is denoted by $x^T y$ or $\langle x, y\rangle$. We use $\mathcal{M}$ to denote a Riemannian manifold. For $x \in \mathcal{M}$, $T_x\mathcal{M}$ denotes the tangent space of $\mathcal{M}$ at x. The inner product defined on $T_x\mathcal{M}$ is denoted by $\langle\cdot,\cdot\rangle_x$, and when no confusion arises, we will also omit the subscript and only use $\langle\cdot,\cdot\rangle$ for simplicity. The tangent bundle $T\mathcal{M} := \cup_x T_x\mathcal{M}$ consists of all tangent vectors to $\mathcal{M}$.

For a smooth function $f : \mathcal{M} \to \mathbb{R}$, the derivative of f at $x \in \mathcal{M}$, denoted by $\mathrm{D}f(x)$, is an element of the dual space to $T_x\mathcal{M}$ which satisfies $\mathrm{D}f(x)v = \lim_{t\to 0} \frac{f(x+tv)-f(x)}{t} := vf$ for all $v \in T_x\mathcal{M}$, where vf is a tangent vector to $\mathbb{R}$ at $f(x)$. That is, let $f : \mathcal{M} \to \mathbb{R}$ be a smooth mapping, $\mathrm{D}f(x)$ is a mapping from $T_x\mathcal{M}$ to $T_{f(x)}\mathbb{R} \simeq \mathbb{R}$. The gradient of f at x (see [5]), denoted by grad $f(x)$, is defined by

$$\langle \text{grad } f(x), v\rangle = \mathrm{D}f(x)v, \quad \forall\, v \in T_x\mathcal{M}.$$

The concept of retraction has played an important role in both theoretic and computational aspects.

Definition 2.1 ([5, p. 55])**.** A retraction on a manifold $\mathcal{M}$ is a smooth mapping R from the tangent bundle $T\mathcal{M}$ onto $\mathcal{M}$ with the following properties. Let R_x denote the restriction of R to $T_x\mathcal{M}$.

1. $R_x(0_x) = x$, where 0_x denotes the zero element of $T_x\mathcal{M}$.
2. With the canonical identification $T_{0x}(T_x\mathcal{M}) \simeq T_x\mathcal{M}, R_x$ satisfies

$$\mathrm{D}R_x(0_x) = id_{T_x\mathcal{M}}$$

where $id_{T_x\mathcal{M}}$ denotes the identity mapping on $T_x\mathcal{M}$.

In the remainder of this paper, we will omit the subscript $T_x\mathcal{M}$ and use **id** to denote the identity mapping $id_{T_x\mathcal{M}}$. For a retraction R_x, define the composite map

$$f_{R_x} = f \circ R_x : T_x\mathcal{M} \to \mathbb{R}.$$

Then $\mathrm{D}f_{R_x}(0) = \mathrm{D}f(x)$. We use $\mathrm{D}^2 f_{R_x}$ to denote the Hessian of f_{R_x}.

Definition 2.2 ([14, p. 608])**.** We say that f_{R_x} is uniformly convex on the $f(x_0)$-sublevel set of f, if there exists $0 < m < M < \infty$ such that

$$m\|v\|^2 \le \mathrm{D}^2 f_{R_x}(p)(v,v) \le M\|v\|^2, \;\; \forall\, v \in T_x\mathcal{M} \tag{2.1}$$

for all $p \in R_x^{-1}(\{\tilde{x} \in \mathcal{M} : f(\tilde{x}) \le f(x_0)\})$.

Definition 2.3 ([5, Chapter 8])**.** We will consider the transport of a vector from one tangent space $T_x\mathcal{M}$ into another one $T_y\mathcal{M}$, that is, consider isomorphisms $\mathcal{T}_{x,y} : T_x\mathcal{M} \longrightarrow T_y\mathcal{M}$. For a retraction R_x, the vector transport $\mathcal{T}^{R_x}_{x,y}$ is defined by

$$\mathcal{T}^{R_x}_{x,y} u := \mathrm{D}R_x(v)[u], \qquad \forall\, u \in T_x\mathcal{M},$$

i.e.

$$\mathcal{T}_{x,y}^{R_x} u = \frac{\mathrm{d}}{\mathrm{d}t} R_x(v+tu)\big|_{t=0}, \qquad \forall\, u \in T_x\mathcal{M},$$

where $v = R_x^{-1}(y)$.

If no confusion, we omit the subscript R_x and use $\mathcal{T}_{x,y}$ to denote $\mathcal{T}_{x,y}^{R_x}$.

For $x \in \mathcal{M}, v \in T_x\mathcal{M}$, assume that R_y^{-1} exists, where $y = R_x(v)$. Then $f_{R_x} = f_{R_y} \circ R_y^{-1} \circ R_x$, and

$$\mathrm{D}f_{R_x}(v) = \mathrm{D}f_{R_y}(0)\mathrm{D}R_x(v) = \mathrm{D}f(y)\mathcal{T}_{x,y}, \tag{2.2}$$

$$\mathrm{grad}\ f_{R_x}(v) = \mathcal{T}_{x,y}^{*}\mathrm{grad}\ f(y), \tag{2.3}$$

where $\mathcal{T}_{x,y}^{*}$ is the adjoint of $\mathcal{T}_{x,y}$ (defined by $\langle u, \mathcal{T}_{x,y}v\rangle = \langle \mathcal{T}_{x,y}^{*}u, v\rangle$ for all $v \in T_x\mathcal{M}$, $u \in T_y\mathcal{M}$).

2.1 Wolfe conditions and BFGS scheme

Given $x \in \mathcal{M}$, for $p \in T_x\mathcal{M}$, if $\langle p, \mathrm{grad}\ f(x)\rangle < 0$, we say that p is a descent direction of f at x.

Definition 2.4 ((Wolfe conditions) [14, p. 600])**.** If the following conditions hold

$$f(R_x(\alpha p)) \le f(x) + \alpha b_1 \mathrm{D}f(x)p, \tag{2.4}$$

$$\mathrm{D}f(R_x(\alpha p))\mathcal{T}_{x,R_x(\alpha p)}p \ge b_2 \mathrm{D}f(x)p, \tag{2.5}$$

where $0 < b_1 < b_2 < 1, 0 < \alpha \le 1$, we say that α satisfies the Wolfe conditions. Note that the above conditions are equivalent to

$$f(R_x(\alpha p)) \le f(x) + \alpha b_1 \mathrm{D}f_{R_x}(0)p, \tag{2.6}$$

$$\langle \mathrm{grad}\ f(R_x(\alpha p)), \mathcal{T}_{x,R_x(\alpha p)}p\rangle \ge b_2\langle \mathrm{grad}\ f(x), p\rangle. \tag{2.7}$$

Replacing (2.7) by

$$|\langle \mathrm{grad}\ f(R_x(\alpha p)), \mathcal{T}_{x,R_x(\alpha p)}p\rangle| \le -b_2\langle \mathrm{grad}\ f(x), p\rangle, \tag{2.8}$$

we obtain the Strong Wolfe conditions.

Assume that x_k is the current iterate and $p_k \in T_{x_k}\mathcal{M}$. Let $x_{k+1} = R_{x_k}(\alpha_k p_k)$, where $\alpha_k > 0$. Define

$$\hat{s}_k := \alpha_k p_k = R_{x_k}^{-1}(x_{k+1}). \tag{2.9}$$

Let $\mathcal{T}_{x_k,x_{k+1}}$ be the vector transport from $T_{x_k}\mathcal{M}$ to $T_{x_{k+1}}\mathcal{M}$, and let

$$s_k \ := \ \mathcal{T}_{x_k,x_{k+1}}\hat{s}_k \in T_{x_{k+1}}\mathcal{M}, \tag{2.10}$$

$$y_k \ := \ \mathrm{grad}\ f(x_{k+1}) - \mathcal{T}_{x_k,x_{k+1}}\mathrm{grad}\ f(x_k) \in T_{x_{k+1}}\mathcal{M}. \tag{2.11}$$

Then the generalization of the secant condition on $\mathcal{M}$ endowed with a vector transport $\mathcal{T}$ is

$$B_{k+1}s_k = y_k, \tag{2.12}$$

where the operator $B_{k+1} : T_{x_{k+1}}\mathcal{M} \longmapsto T_{x_{k+1}}\mathcal{M}$. The BFGS scheme on $\mathcal{M}$ is as follows

$$B_{k+1}p = \hat{B}_k p - \frac{\langle s_k, \hat{B}_k p\rangle}{\langle s_k, \hat{B}_k s_k\rangle}\hat{B}_k s_k + \frac{\langle y_k, p\rangle}{\langle y_k, s_k\rangle}y_k, \quad \forall\, p \in T_{x_{k+1}}\mathcal{M}, \tag{2.13}$$

with $\hat{B}_k = \mathcal{T}_{x_k,x_{k+1}} \circ B_k \circ \mathcal{T}_{x_k,x_{k+1}}^{-1}$, see [5].

3 Subspace Method on Riemannian Manifolds

In this section, we generalize Yuan and Stoer's subspace method from $\mathbb{R}^n$ to a Riemannian manifold $\mathcal{M}$. Now we take some effort to describe Yuan and Stoer's subspace method briefly. For detail, the reader can refer to [20].

3.1 Yuan and Stoer's subspace method

For a twice continuously differentiable function f defined on $\mathbb{R}^n$, the quadratic approximate of f at iterate x_{k+1} is

$$f(x) \approx f(x_{k+1}) + g_{k+1}^T(x - x_{k+1}) + \frac{1}{2}(x - x_{k+1})^T B_{k+1}(x - x_{k+1}),$$

where g_{k+1} is the gradient of f at x_{k+1} and B_{k+1} is an approximation to the Hessian $\nabla^2 f(x_{k+1})$. Then, to get the descent direction p_{k+1}, the most general method is to minimize $\varphi_{k+1}(p)$ subject to $p \in \mathbb{R}^n$. However, Yuan and Stoer consider the following problem

$$\min_{p\in\Omega_k} \varphi_{k+1}(p), \tag{3.1}$$

where $\Omega_k = \mathrm{span}\{g_{k+1}, p_k\}$ and p_k is the search direction at x_k.

Assume that B_{k+1} satisfies the secant equation $B_{k+1}s_k = y_k$. The reader can refer to [20] about the definition of s_k, y_k. Substituting p by $\mu g_{k+1} + \nu s_k$ in (3.1), we obtain that

$$\min_{(\mu,\nu)\in\mathbb{R}^2} \begin{pmatrix} \|g_{k+1}\|^2 \\ \langle g_{k+1}, s_k\rangle \end{pmatrix}^T \begin{pmatrix} \mu \\ \nu \end{pmatrix} + \frac{1}{2}(\mu,\nu) \begin{pmatrix} \rho_k & \langle g_{k+1}, y_k\rangle \\ \langle y_k, g_{k+1}\rangle & \langle y_k, s_k\rangle \end{pmatrix} \begin{pmatrix} \mu \\ \nu \end{pmatrix} \tag{3.2}$$

where $\rho_k = \langle B_{k+1}g_{k+1}, g_{k+1}\rangle$. This method has several advantages: Firstly, the solution p_{k+1} of (3.2) can be easily computed. Secondly, p_{k+1} obtains the optimal decrease in the subspace $\mathrm{span}\{g_{k+1}, p_k\}$, while the search direction of the nonlinear conjugate gradient method usually does not. Therefore Yuan and Stoer's method is at least as effective as the nonlinear conjugate gradient method.

3.2 Generalization of Yuan and Stoer's method

When generalizing Yuan and Stoer's method from $\mathbb{R}^n$ to $\mathcal{M}$, the main difficulty is the following: since $p_k \in T_{x_k}\mathcal{M}$ and $\mathrm{grad}\ f(x_{k+1})$ belongs to another tangent space $T_{x_{k+1}}\mathcal{M}$, the situation results in the nonexistence of $\mathrm{span}\{\mathrm{grad}\ f(x_{k+1}), p_k\}$. Fortunately, the strategy of transporting p_k from $T_{x_k}\mathcal{M}$ to $T_{x_{k+1}}\mathcal{M}$ can remedy this problem. Consider the quadratic approximation of f at x_{k+1}:

$$\min_{p\in T_{x_{k+1}}\mathcal{M}} \hat{m}_{x_{k+1}}(p) = f(x_{k+1}) + \langle \mathrm{grad}\ f(x_{k+1}), p\rangle + \frac{1}{2}\langle B_{k+1}p, p\rangle, \tag{3.3}$$

where B_{k+1} is unknown. But it has to satisfy the secant condition (2.12). Let p_k be the search direction at x_k. Then $p_k \in T_{x_k}\mathcal{M}$.

To generalize Yuan and Stoer's method, we need to transport p_k to the space $T_{x_{k+1}}\mathcal{M}$. Define $\Omega_k := \mathrm{span}\{\mathrm{grad}\ f(x_{k+1}), \mathcal{T}_{x_k,x_{k+1}}p_k\}$. Then we obtain a minimization problem on a two dimensional subspace Ω_k:

$$\min_{p\in\Omega_k} \hat{m}_{x_{k+1}}(p).$$

In the remainder of this paper, we use the notation

$$g_k = \text{grad } f(x_k), \qquad \forall k \geq 0$$

Since $s_k = \mathcal{T}_{x_k,x_{k+1}} \hat{s}_k = \alpha_k \mathcal{T}_{x_k,x_{k+1}} p_k$, if $p \in \Omega_k$, then $p = \mu g_{k+1} + \nu s_k$ for some $\mu, \nu \in \mathbb{R}$. Substituting it into (3.3), we obtain a function

$$\psi(\mu, \nu) := \hat{m}_{x_{k+1}}(\mu g_{k+1} + \nu s_k), \tag{3.4}$$

which is just (3.2) except that the inner product $\langle \cdot, \cdot \rangle$ is defined on $T_{x_{k+1}}\mathcal{M}$ and a constant term $f(x_{k+1})$.

As in [20], we consider separately the two cases: (1) g_{k+1} and s_k are collinear; (2) g_{k+1} and s_k are not collinear.

For the first case, if there exists a λ such that $g_{k+1} = \lambda s_k$, then as in [20, (2.8)], the next search direction is set to be

$$p_{k+1} = -\frac{\langle g_{k+1}, s_k \rangle}{\langle y_k, s_k \rangle} s_k. \tag{3.5}$$

For the second case, assume that ρ_k satisfies the relation

$$\rho_k \langle y_k, s_k \rangle - \langle g_{k+1}, y_k \rangle^2 > 0, \tag{3.6}$$

the unique solution of $\psi(\mu, \nu)$ is

$$\begin{pmatrix} \mu_{k+1} \\ \nu_{k+1} \end{pmatrix} = \frac{-1}{\rho_k \langle y_k, s_k \rangle - \langle g_{k+1}, y_k \rangle^2} \begin{pmatrix} \langle y_k, s_k \rangle \|g_{k+1}\|^2 - \langle g_{k+1}, y_k \rangle \langle g_{k+1}, s_k \rangle \\ \rho_k \langle g_{k+1}, s_k \rangle - \langle g_{k+1}, y_k \rangle \|g_{k+1}\|^2 \end{pmatrix}.$$

Thus, the search direction p_{k+1} can be chosen as

$$\begin{aligned} p_{k+1} =& \mu_{k+1} g_{k+1} + \nu_{k+1} s_k \\ =& \frac{1}{\rho_k \langle y_k, s_k \rangle - \langle g_{k+1}, y_k \rangle^2} \big[(\langle g_{k+1}, y_k \rangle \langle g_{k+1}, s_k \rangle - \langle y_k, s_k \rangle \|g_{k+1}\|^2) g_{k+1} \\ & + (\langle g_{k+1}, y_k \rangle \|g_{k+1}\|^2 - \rho_k \langle g_{k+1}, s_k \rangle) s_k \big]. \end{aligned} \tag{3.7}$$

Note that the above formula has been derived in [20], we only give it for completeness.

Of course, different values of ρ_k give different p_{k+1}. There are two choices of ρ_k supplied by Yuan and Stoer: one is

$$\rho_k = \frac{\langle y_k, s_k \rangle}{\|s_k\|^2} \Big(\|g_{k+1}\|^2 - \frac{\langle g_{k+1}, s_k \rangle^2}{\|s_k\|^2} \Big) + \frac{\langle g_{k+1}, y_k \rangle^2}{\langle y_k, s_k \rangle}, \tag{3.8}$$

which is obtained from $\rho_k = \langle B_{k+1} g_{k+1}, g_{k+1} \rangle$, where

$$B_{k+1} p = \frac{\langle y_k, s_k \rangle}{\|s_k\|^2} \big(p - \frac{\langle s_k, p \rangle}{\|s_k\|^2} s_k \big) + \frac{\langle y_k, p \rangle}{\langle y_k, s_k \rangle} y_k. \tag{3.9}$$

corresponding to (2.13) when $\hat{B}_k = \frac{\langle y_k, s_k \rangle}{\|s_k\|^2} id$.

The other is

$$\rho_k = 2 \frac{\langle g_{k+1}, y_k \rangle^2}{\langle y_k, s_k \rangle}, \tag{3.10}$$

which is based on the interval of ρ_k.

Now we state the overall algorithm.

Algorithm: Subspace quasi-Newton optimization method on Riemannian manifolds

---

Require: Riemannian manifold $\mathcal{M}$; vector transport $\mathcal{T}$ on $\mathcal{M}$ with associated retraction R; real-valued function f on $\mathcal{M}$.
Goal: Find a minimizer of f.
Parameters: $\epsilon \geq 0$, $0 < b_1 < b_2 < 1$.
Input: Initial iterate $x_0 \in \mathcal{M}$.
Output: Point x^* such that $\|\text{grad} f(x^*)\| \leq \epsilon$.
Step 1: k=0; set $p_0 = -\text{grad } f(x_0)$;
Step 2: Compute a step length α_k satisfying the strong Wolfe conditions;
set $x_{k+1} = R_{x_k}(\alpha_k p_k)$;
compute $g_{k+1} = \text{grad } f(x_{k+1})$;
if $\|g_{k+1}\| \leq \epsilon$ then stop.
else go to step 3;
Step 3: If $\langle g_{k+1}, \mathcal{T}_{x_k,x_{k+1}} p_k\rangle$ are collinear,
then define p_{k+1} by (3.5) and go to step 5;
else go to step 4;
Step 4: Choose ρ_k satisfying (3.6) ;
compute p_{k+1} by (3.7) ;
Step 5: k:=k+1, go to Step 2.

---

4 Convergence Analysis

In this section, we show that our algorithm is globally convergent. Under some conditions, the local linear convergence of our method can also be established. To prove the convergence results, we always assume that:
Assumption A: f is twice continuously differentiable and bounded below.
Assumption B: $\mathcal{T}_{x_k,x_{k+1}}$ is an isometry for all $k \geq 1$, where x_k is the iterate generated by our subspace algorithm.

Given a descent direction p, the following result tells us that a step length satisfying the Wolfe conditions always exists.

Lemma 4.1 (Feasible step length, e.g. [12])**.** *If $p \in T_x\mathcal{M}$ is a descent direction at $x \in \mathcal{M}$, then there exists $\alpha > 0$ that satisfies Wolfe conditions* (2.6) *and* (2.7)*.*

In the following lemma, we prove that if α_k satisfies the Wolfe conditions, then p_k is always a descent direction at x_k for all $k \geq 1$.

Lemma 4.2. *Assume that $\mathcal{T}_{x_k,x_{k+1}}$ is an isometry. If p_k is a descent direction at x_k and α_k satisfies the Wolfe conditions, then $\langle y_k, s_k\rangle > 0$ and p_{k+1} is a descent direction at x_{k+1}.*

Proof. From (2.10), (2.11), and the Wolfe condition (2.7), it follows that

$$\begin{aligned}
\langle y_k, s_k\rangle &= \langle g_{k+1} - \mathcal{T}_{x_k,x_{k+1}} g_k, \mathcal{T}_{x_k,x_{k+1}} \hat{s}_k\rangle \\
&= \langle g_{k+1}, \mathcal{T}_{x_k,x_{k+1}} \hat{s}_k\rangle - \langle g_k, \hat{s}_k\rangle \\
&\geq b_2\langle g_k, \hat{s}_k\rangle - \langle g_k, \hat{s}_k\rangle \\
&= \alpha_k(b_2 - 1)\langle g_k, p_k\rangle > 0, \qquad (4.1)
\end{aligned}$$

where the last inequality follows from $b_2 < 1$ and the assumption that p_k is a descent direction at x_k.

If g_{k+1} and s_k are collinear, from (3.5) and (4.1), it follows that $\langle g_{k+1}, p_{k+1}\rangle < 0$. Now assume that g_{k+1} and s_k are not collinear. Since (μ_{k+1}, ν_{k+1}) is the optimal solution of $\psi(\mu, \nu)$ defined by (3.4), we have

$$\begin{aligned} -\langle g_{k+1}, p_{k+1}\rangle =& 2[\psi(0,0) - \psi(\mu_{k+1}, \nu_{k+1})] \\ \geq& 2\Big[\psi(0,0) - \psi(-\frac{\|g_{k+1}\|^2}{\rho_k}, 0)\Big] \\ =& \frac{\|g_{k+1}\|^4}{\rho_k}. \end{aligned} \tag{4.2}$$

Whatever ρ_k is defined by (3.8) or (3.10), we have $\rho_k > 0$, which together with (4.2) implies the second assertion. □

Note that throughout this subsection $\hat{s}_k$, s_k and y_k are defined by (2.9), (2.10) and (2.11). The following theorem treats the case ρ_k is defined by (3.10).

Theorem 4.3. *Choosing ρ_k by (3.10). Assume that $f_{R_{x_k}}$ is uniformly convex on the $f(x_0)$-sublevel set of f. If there exist $\hat{M}, \hat{\delta} > 0$ such that*

$$\hat{\delta}\min\{1, \|g_{k+1}\|\}\langle y_k, s_k\rangle \leq \rho_k\langle y_k, s_k\rangle - \langle g_{k+1}, y_k\rangle^2 \leq \hat{M}\langle y_k, s_k\rangle, \quad \forall\ k, \tag{4.3}$$

then

$$\liminf_{k\to\infty} \|\text{grad } f(x_k)\| = 0.$$

Proof. If not, there is $\delta > 0$ such that

$$\|g_k\| \geq \delta, \qquad \forall k \geq 0 \tag{4.4}$$

By the Wolfe conditon (2.6), for any k, we have

$$f(x_{k+1}) \leq f(x_k) + b_1 \mathrm{D} f_{R_{x_k}}(0)(\hat{s}_k) = f(x_k) + b_1\langle g_k, \hat{s}_k\rangle.$$

Since $\{f(x_k)\}$ is a non-increasing sequence and f is bounded below, it follows from the above inequality that

$$\sum_{k=1}^{\infty} -\langle g_k, \hat{s}_k\rangle < +\infty. \tag{4.5}$$

From (4.1), it follows that

$$\langle y_k, s_k\rangle \geq (b_2 - 1)\langle g_k, \hat{s}_k\rangle. \tag{4.6}$$

By Lemma 4.2, we have $\langle y_k, s_k\rangle > 0$ and $\langle g_k, \hat{s}_k\rangle < 0$. Therefore, the inequality (4.6) becomes

$$(b_2 - 1)\frac{\langle g_k, \hat{s}_k\rangle}{\langle y_k, s_k\rangle} \leq 1.$$

Multiplying it by $-\langle g_k, \hat{s}_k\rangle$ on two sides, we obtain

$$(1 - b_2)\frac{\langle g_k, \hat{s}_k\rangle^2}{\langle y_k, s_k\rangle} \leq -\langle g_k, \hat{s}_k\rangle,$$

which, along with (4.5), yields

$$\sum_{k=1}^{\infty} \frac{\langle g_k, \hat{s}_k\rangle^2}{\langle y_k, s_k\rangle} < +\infty. \tag{4.7}$$

From (2.2), it follows that

$$\langle g_{k+1}, s_k\rangle = \mathrm{D}f(x_{k+1})s_k = \mathrm{D}f_{R_{x_k}}(\hat{s}_k)\hat{s}_k. \tag{4.8}$$

By (2.10) and (2.11), we have

$$\langle y_k, s_k\rangle = (\mathrm{D}f_{R_{x_k}}(\hat{s}_k) - \mathrm{D}f_{R_{x_k}}(0))\hat{s}_k = \mathrm{D}^2 f_{R_{x_k}}(\theta\hat{s}_k)(\hat{s}_k, \hat{s}_k),$$

in which $\theta \in [0, 1]$. By (4.2) and (2.9), we have $\langle g_k, \theta\hat{s}_k\rangle = \alpha_k\theta\langle g_k, p_k\rangle \leq -\alpha_k\theta\frac{\|g_k\|^4}{\rho_{k-1}} < 0$, which yields that $f(R_{x_k}(\theta\hat{s}_k)) < f(x_0)$. Since $f_{R_{x_k}}$ is uniformly convex on the $f(x_0)$-sublevel set of f, we have

$$m\|\hat{s}_k\|^2 \leq \langle y_k, s_k\rangle \leq M\|\hat{s}_k\|^2,$$

which implies that

$$\langle y_k, s_k\rangle = O(\|s_k\|^2) = O(\|\hat{s}_k\|^2). \tag{4.9}$$

By (4.7) and (4.9), we have

$$\sum_{k=1}^{\infty} \frac{\langle g_k, p_k\rangle^2}{\|p_k\|^2} < +\infty. \tag{4.10}$$

Substituting (4.8) into (2.8), we can get $|\langle g_{k+1}, s_k\rangle| \leq -b_2\langle g_k, \hat{s}_k\rangle$, which together with (4.6) yields

$$|\langle g_{k+1}, s_k\rangle| \leq \frac{b_2}{1-b_2}\langle y_k, s_k\rangle. \tag{4.11}$$

Similar as in [14, p. 620], we can define the averaged Hessian G_k and $\hat{y}_k$ as

$$G_k := \int_0^1 \mathrm{D}^2 f_{R_{x_k}}(t\hat{s}_k)\mathrm{d}t, \quad \hat{y}_k := \mathrm{D}f_{R_{x_k}}(\hat{s}_k) - \mathrm{D}f_{R_{x_k}}(0).$$

Then $\langle y_k, s_k\rangle = \hat{y}_k\hat{s}_k$, $G_k(\hat{s}_k, \cdot) = \hat{y}_k$ and in addition to (2.2), (2.3), we get

$$\begin{aligned}
\|\hat{y}_k\| &= \max_{v\in T_{x_k}\mathcal{M}} \frac{(\mathrm{D}f_{R_{x_k}}(\hat{s}_k) - \mathrm{D}f_{R_{x_k}}(0))v}{\|v\|} \\
&= \max_{v\in T_{x_k}\mathcal{M}} \frac{\mathrm{D}f_{R_{x_k}}(\hat{s}_k)v - \mathrm{D}f_{R_{x_k}}(0)v}{\|v\|} \\
&= \max_{v\in T_{x_k}\mathcal{M}} \frac{\langle g_{k+1}, \mathcal{T}_{x_k,x_{k+1}}v\rangle - \langle g_k, v\rangle}{\|v\|} \\
&= \max_{v\in T_{x_k}\mathcal{M}} \frac{\langle g_{k+1}, \mathcal{T}_{x_k,x_{k+1}}v\rangle - \langle \mathcal{T}_{x_k,x_{k+1}}g_k, \mathcal{T}_{x_k,x_{k+1}}v\rangle}{\mathcal{T}_{x_k,x_{k+1}}v} \\
&= \max_{\mathcal{T}_{x_k,x_{k+1}}v\in T_{x_{k+1}}\mathcal{M}} \frac{\langle y_k, \mathcal{T}_{x_k,x_{k+1}}v\rangle}{\|\mathcal{T}_{x_k,x_{k+1}}v\|} \\
&= \|y_k\|.
\end{aligned}$$

Let $\hat{G}_k$ be the Lax-Milgram representation of G_k. Then we have

$$\frac{\|y_k\|^2}{\langle y_k, s_k\rangle} = \frac{\|\hat{y}_k\|^2}{\hat{y}_k\hat{s}_k} = \frac{G_k(\sqrt{\hat{G}_k}\hat{s}_k, \sqrt{\hat{G}_k}\hat{s}_k)}{\|\sqrt{\hat{G}_k}\hat{s}_k\|^2} \leq M. \tag{4.12}$$

From (3.10) and (4.3), it follows that

$$2\hat{\delta} \le \rho_k \le 2\hat{M}. \tag{4.13}$$

By (3.7) and (4.3), we have

$$\|p_{k+1}\| \le \frac{1}{\hat{\delta}\min\{1, \|g_{k+1}\|\}\langle y_k, s_k\rangle}\big[3\|y_k\|\|s_k\|\|g_{k+1}\|^3 + \rho_k|\langle g_{k+1}, s_k\rangle|\|s_k\|\big]. \tag{4.14}$$

By (4.5), $\langle g_k, p_k\rangle$ is bounded, which together with $\alpha_k \le 1$ implies that

$$\|\hat{s}_k\| = \alpha_k\|p_k\| = O\big(\frac{\|p_k\|}{-\langle g_k, p_k\rangle}\big). \tag{4.15}$$

With the above inequalities in hand, now we prove that the sequence $\{\|p_k\|/(-\langle g_k, p_k\rangle)\}$ is decreasing. By (4.2), we know that

$$\frac{\|p_{k+1}\|}{-\langle g_{k+1}, p_{k+1}\rangle} \le \frac{\rho_k\|p_{k+1}\|}{\|g_{k+1}\|^4},$$

which together with (4.4) and (4.14) implies that

$$\begin{aligned}
\frac{\|p_{k+1}\|}{-\langle g_{k+1}, p_{k+1}\rangle} \le& \frac{\rho_k}{\langle y_k, s_k\rangle} \cdot \frac{1}{\hat{\delta}\min\{1, \|g_{k+1}\|\}}\big[3\frac{\|y_k\|\|s_k\|}{\|g_{k+1}\|} + \rho_k\frac{|\langle g_{k+1}, s_k\rangle|\|s_k\|}{\|g_{k+1}\|^4}\big] \\
\le& \frac{\rho_k}{\langle y_k, s_k\rangle}[O(\|y_k\|\|s_k\|) + O(\rho_k \cdot |\langle g_{k+1}, s_k\rangle| \cdot \|s_k\|)] \\
\le& \frac{\rho_k}{\langle y_k, s_k\rangle}[O(\sqrt{M\langle y_k, s_k\rangle}\|s_k\|) + O(\rho_k \cdot \frac{b_2}{1-b_2}\langle y_k, s_k\rangle \cdot \|s_k\|)] (\text{by } (4.11), (4.12)) \\
\le& O(\frac{\|s_k\|}{\sqrt{\langle y_k, s_k\rangle}}) + O(\|s_k\|) \quad\quad (\text{by } (4.13)) \\
=& O(\frac{\|\hat{s}_k\|}{\sqrt{\langle y_k, s_k\rangle}}) + O(\|\hat{s}_k\|) \\
=& O(\sqrt{\|\hat{s}_k\|} \cdot \frac{\|\hat{s}_k\|}{\sqrt{\langle y_k, s_k\rangle}}) + O(\|\hat{s}_k\|) \\
\le& O\Big(\sqrt{\|\hat{s}_k\|} \cdot \sqrt{\frac{\|\hat{s}_k\|}{-(1-b_2)\langle g_k, \hat{s}_k\rangle}}\Big) + O(\|\hat{s}_k\|) \quad\quad (\text{by } (4.6)) \\
\le& O\big(\sqrt{\|\hat{s}_k\|}\sqrt{\frac{\|p_k\|}{-\langle g_k, p_k\rangle}}\big) + O(\sqrt{\|\hat{s}_k\|} \cdot \sqrt{\|\hat{s}_k\|}) \\
\le& O\big(\sqrt{\|\hat{s}_k\|}\sqrt{\frac{\|p_k\|}{-\langle g_k, p_k\rangle}}\big) \quad\quad (\text{by } (4.15)) \\
\le& O\big(\sqrt{-\langle g_k, \hat{s}_k\rangle}\frac{\|p_k\|}{-\langle g_k, p_k\rangle}\big).
\end{aligned} \tag{4.16}$$

Note that by (4.5), the term $\sqrt{-\langle g_k, \hat{s}_k\rangle}$ in (4.16) tends to zero as k goes to infinity. Thus $\frac{\|p_{k+1}\|}{-\langle g_{k+1}, p_{k+1}\rangle} \le \frac{\|p_k\|}{-\langle g_k, p_k\rangle}$ for all sufficiently large k. Therefore $\frac{\langle g_k, p_k\rangle^2}{\|p_k\|^2} \ge \tau$ for some $\tau > 0$, which contradicts (4.10). The proof is complete. $\square$

Now we study the case that ρ_k is chosen from (3.8). The following result will be useful in our analysis.

Lemma 4.4. *Assume that $f_{R_{x_k}}$ is uniformly convex on the $f(x_0)$-sublevel set of f. Let B_k be defined by (3.9). Then B_k is positive definite. Moreover, $\|B_k\|$ and $\|B_k^{-1}\|$ is uniformly bounded.*

Proof. Note that B_{k+1} is just the one step RBFGS update from $\frac{\langle y_k, s_k\rangle}{\|s_k\|^2} id$. By (4.9), there exist $m, M > 0$ such that $m < \frac{\langle y_k, s_k\rangle}{\|s_k\|^2} < M$. For any unit $p(\|p\| = 1)$, we have

$$\begin{aligned}\langle B_{k+1}p, p\rangle &= \frac{\langle y_k, s_k\rangle}{\|s_k\|^2}(\langle p, p\rangle - \frac{\langle s_k, p\rangle^2}{\|s_k\|^2}) + \frac{\langle y_k, p\rangle^2}{\langle y_k, s_k\rangle} \\ &= \frac{\langle y_k, s_k\rangle}{\|s_k\|^2}(1 - \frac{\langle s_k, p\rangle^2}{\|s_k\|^2}) + \frac{\langle y_k, p\rangle^2}{\langle y_k, s_k\rangle}.\end{aligned}$$

Since $0 < \frac{\langle s_k, p\rangle^2}{\|s_k\|^2} \le 1$, we have

$$0 < \frac{\langle y_k, p\rangle^2}{\langle y_k, s_k\rangle} \le \langle B_{k+1}p, p\rangle \le \frac{\langle y_k, s_k\rangle}{\|s_k\|^2} + \frac{\langle y_k, p\rangle^2}{\langle y_k, s_k\rangle} \le \frac{\langle y_k, s_k\rangle}{\|s_k\|^2} + \frac{\|y_k\|^2}{\langle y_k, s_k\rangle} \le 2M,$$

in which the last inequality follows from (4.12). Furthermore, $\|B_{k+1}\| \ge \frac{\|y_k\|^2}{\langle y_k, s_k\rangle} \ge m$. Then $\|B_{k+1}\|$ and $\|B_{k+1}^{-1}\|$ are positive definite and uniformly bounded. □

Let $f^* := \min_{x\in\mathcal{M}} f(x)$. The following theorem, which tells us $f(x_k) - f^*$ converges to zero linearly, is another main result of this subsection.

Theorem 4.5. *Choosing ρ_k by (3.8). Assume that $f_{R_{x_k}}$ is uniformly convex on the $f(x_0)$-sublevel set of f, the sequence $\{x_k\}$ is formed by the Riemannian subspace quasi-Newton algorithm, there exists a constant $\mu \in (0, 1)$ such that*

$$f(x_k) - f^* \le \mu^k(f(x_0) - f^*).$$

Proof. Let B_k be defined by (3.9). Define θ_k and q_k by

$$\theta_k = \arccos \frac{\langle B_k\hat{s}_k, \hat{s}_k\rangle}{\|\hat{s}_k\|\|B_k\hat{s}_k\|}, \quad q_k = \frac{\langle B_k\hat{s}_k, \hat{s}_k\rangle}{\|\hat{s}_k\|^2} = \frac{\langle B_kp_k, p_k\rangle}{\|p_k\|^2}. \tag{4.17}$$

From (2.7), it follows that $-\mathrm{D}f_{R_{x_k}}(\alpha_kp_k)p_k \le -b_2\mathrm{D}f(x_k)p_k$, which together with

$$\begin{aligned}-\mathrm{D}f_{R_{x_k}}(\alpha_kp_k)p_k &= -\mathrm{D}f(x_k)p_k - \alpha_k\int_0^1 \mathrm{D}^2f_{R_{x_k}}(t\alpha_kp_k)(p_k, p_k)\mathrm{dt} \\ &\ge -\mathrm{D}f(x_k)p_k - \alpha_kM\|p_k\|^2\end{aligned}$$

implies that $-b_2\mathrm{D}f(x_k)p_k \ge -\mathrm{D}f(x_k)p_k - \alpha_kM\|p_k\|^2$. Thus,

$$\alpha_k \ge \frac{b_2 - 1}{M} \cdot \frac{\mathrm{D}f(x_k)p_k}{\|p_k\|^2} = \frac{b_2 - 1}{M} \cdot \frac{\langle g_k, p_k\rangle}{\|p_k\|^2}.$$

By our subspace algorithm, we have the relation $\langle B_kp_k, p_k\rangle = -\langle g_k, p_k\rangle$. It follows from (4.17) that

$$\alpha_k \ge \frac{1 - b_2}{M} \cdot \frac{\langle B_kp_k, p_k\rangle}{\|p_k\|^2} = \frac{1 - b_2}{M}q_k.$$

By (2.4), (4.2) and Lemma 4.4, we have

$$\begin{aligned} f(x_k) - f(x_{k+1}) &\geq -\alpha_k b_1 \mathrm{D} f(x_k) p_k \\ &\geq \alpha_k b_1 \frac{\|g_k\|^4}{\rho_{k-1}} = \alpha_k b_1 \frac{\|g_k\|^2}{\langle B_k g_k, g_k\rangle} \|g_k\|^2 \\ &\geq b_1 \frac{1-b_2}{M^2} \frac{q_k}{\cos^2\theta_k} \cos^2\theta_k \|g_k\|^2. \end{aligned} \tag{4.18}$$

Now we prove that there exists $\beta > 0$ such that $\cos\theta_k \geq \beta$ for all k. Since B_k is defined by (3.9), it is one-step RBFGS update from $\Lambda = \lambda\ id$, where $\lambda = \frac{\langle y_{k-1}, s_{k-1}\rangle}{\|s_{k-1}\|^2}$. Let $\bar{B}_{k+1}$ be the BFGS update of B_k. Then

$$\begin{aligned} \mathrm{tr}(\frac{1}{\lambda}\bar{B}_{k+1} - id) &= \mathrm{tr}(\frac{1}{\lambda}\big[B_k - \frac{\langle B_k \hat{s}_k, \cdot\rangle B_k \hat{s}_k}{\langle B_k \hat{s}_k, \hat{s}_k\rangle} + \frac{\hat{y}_k(\cdot) y_k}{\langle y_k, s_k\rangle}\big] - id) \\ &= \mathrm{tr}(\frac{1}{\lambda}B_k - id) - \frac{\|B_k \hat{s}_k\|^2}{\lambda\langle B_k \hat{s}_k, \hat{s}_k\rangle} + \frac{\|y_k\|^2}{\lambda\langle y_k, s_k\rangle} \\ &= \mathrm{tr}(\frac{1}{\lambda}B_k - id) - \frac{\langle B_k \hat{s}_k, \hat{s}_k\rangle}{\lambda\|\hat{s}_k\|^2} \frac{\|B_k \hat{s}_k\|^2 \|\hat{s}_k\|^2}{\langle B_k \hat{s}_k, \hat{s}_k\rangle^2} + \frac{\|y_k\|^2}{\lambda\langle y_k, s_k\rangle} \\ &\leq \mathrm{tr}(\frac{1}{\lambda}B_k - id) - \frac{q_k}{\lambda\cos^2\theta_k} + \frac{M}{\lambda} \\ &\leq \mathrm{tr}(\frac{1}{\lambda}\Lambda - id) - \frac{\|\Lambda \hat{s}_{k-1}\|^2}{\lambda\langle \Lambda \hat{s}_{k-1}, \hat{s}_{k-1}\rangle} + \frac{\|y_{k-1}\|^2}{\lambda\langle y_{k-1}, s_{k-1}\rangle} - \frac{q_k}{\lambda\cos^2\theta_k} + \frac{M}{\lambda} \\ &\leq -1 + \frac{M}{\lambda} - \frac{q_k}{\lambda\cos^2\theta_k} + \frac{M}{\lambda} \\ &\leq \frac{2M}{\lambda} - \frac{q_k}{\lambda\cos^2\theta_k} - 1. \end{aligned} \tag{4.19}$$

By the proof of [14, p.621], we have

$$\det(\frac{1}{\lambda}\bar{B}_{k+1}) \geq \frac{m}{q_k}\det(\frac{1}{\lambda}B_k) \geq \frac{m}{q_k}\frac{m\|\hat{s}_{k-1}\|^2}{\langle \Lambda \hat{s}_{k-1}, \hat{s}_{k-1}\rangle} \geq \frac{m}{\lambda}\frac{m}{q_k}. \tag{4.20}$$

Let $\Phi(B) := \mathrm{tr}(B - \mathrm{id}) - \log\det B$. By Lidskii's theorem (see [15, Thm. 3.5]), if B is positive definite, then $\psi(B) \geq 0$. Combining this inequality with (4.19) and (4.20) yields

$$\begin{aligned} 0 \leq \Phi(\frac{1}{\lambda}\bar{B}_{k+1}) &\leq \frac{2M}{\lambda} - \frac{q_k}{\lambda\cos^2\theta_k} - \log(\frac{m}{\lambda}\cdot\frac{m}{q_k}) - 1 \\ &\leq \frac{2M}{\lambda} - 2\log m - 2 + \log(\lambda^2\cos^2\theta_k) + 1 - \frac{q_k}{\lambda\cos^2\theta_k} + \log\frac{q_k}{\lambda\cos^2\theta_k}. \end{aligned}$$

By (4.9), λ is bounded. Then we have

$$\log(\lambda^2\cos^2\theta_k) + 1 - \frac{q_k}{\lambda\cos^2\theta_k} + \log\frac{q_k}{\lambda\cos^2\theta_k} \geq C$$

for some real number C. Let $g(z) := 1 - z + \log z$, where $z > 0$. Then $g(z) < 0$ for any $z > 0$, which implies

$$1 - \frac{q_k}{\lambda\cos^2\theta_k} + \log\frac{q_k}{\lambda\cos^2\theta_k} < 0.$$

From the above two inequalities, it follows that

$$\log(\lambda^2 \cos^2 \theta_k) \geq C. \tag{4.21}$$

Then there exists $\beta > 0$ such that $\cos\theta_k \geq \beta$ for all k. And there exists κ such that $\frac{q_k}{\cos^2\theta_k} \geq \kappa$.
By (4.18) and $\cos\theta_k \geq \beta$, we have

$$f(x_k) - f(x_{k+1}) \geq b_1 \kappa \beta^2 \frac{1-b_2}{M^2} \|g_k\|^2. \tag{4.22}$$

Note that $y \in \mathcal{M}$ is in the neighborhood of x, there exists $t \in [0,1]$ such that

$$\begin{aligned} f(y) - f(x) &= \mathrm{D}f(x) R_x^{-1}(y) + \frac{1}{2}\mathrm{D}^2 f_{R_x}(tR_x^{-1}(y))(R_x^{-1}(y), R_x^{-1}(y)) \\ &\geq \mathrm{D}f(x)R_x^{-1}(y) + \frac{m}{2}\|R_x^{-1}(y)\|^2 \\ &\geq -\frac{1}{2m}\|\mathrm{D}f(x)\|^2 = -\frac{1}{2m}\|g(x)\|^2. \end{aligned}$$

Since $y \in \mathcal{M}$ is arbitrary, we have $f(x) - f^* \leq \|g(x)\|^2/(2m)$. Combining it with (4.22) yields

$$f(x_k) - f(x_{k+1}) \geq 2mb_1\kappa\beta^2 \frac{1-b_2}{M^2}(f(x_k) - f^*), \tag{4.23}$$

which implies that

$$f(x_k) - f(x^*) \leq \mu^k (f(x_0) - f(x^*)),$$

where $\mu = 1 - 2mb_1\kappa\beta^2\frac{1-b_2}{M^2}$. □

Next we prove the local linear convergence of the subspace algorithm. In the proof of the last result, we adopt the notation $F(x) = \Omega(G(x))$ as $x \to x^*$, which means that there exist $l, L > 0$ and a neighborhood $\mathcal{N}$ of x^* such that $l\|F(x)\| \leq \|G(x)\| \leq L\|F(x)\|$ for all $x \in \mathcal{N}$.

Theorem 4.6. *Suppose that the assumptions of Theorem 4.5 hold. Assume that x_k converges to an optimal solution x^*. There exists a K such that for all $k \geq K$, there exist $\tau > 0$ and $\mu \in (0,1)$ such that*

$$\mathrm{dist}(x_k, x^*) \leq \tau\mu^{k-K}\mathrm{dist}(x_K, x^*).$$

Proof. There exists a neighborhood $\mathcal{U}$ of x^* such that, for all $x \in \mathcal{U}$,

$$f(x) - f(x^*) = \frac{1}{2}\mathrm{D}^2 f_{R_{x^*}}(tR_{x^*}^{-1}(x))(R_{x^*}^{-1}(x), R_{x^*}^{-1}(x))$$

for some $t \in (0,1)$, which together with (2.1) implies that $f(x) - f(x^*) = \Omega(\|R_{x^*}^{-1}(x)\|^2)$. From the proof of [5, Prop. 7.1.3] it follows that $\|R_{x^*}^{-1}(x)\| = \Omega(\mathrm{dist}(x, x^*))$, and therefore

$$f(x) - f(x^*) = \Omega(\mathrm{dist}(x, x^*)^2). \tag{4.24}$$

Since $\{x_k\}$ converges to x^*, there is a K such that, for all $k > K$, x_k belongs to $\mathcal{U}$. By (4.23), we have

$$f(x_k) - f(x^*) \leq \mu(f(x_{k-1}) - f(x^*)) \leq \mu^{k-K}(f(x_K) - f(x^*)).$$

Then the assertion follows from (4.24). □

$\boxed{5}$ Numerical Results

In this section, we demonstrate the effectiveness of our Riemannian subspace quasi-Newton algorithm on some test problems. All of our tests are carried out in MATLAB R2014a on a Thinkpad notebook Intel Core i5 with 2.53GHz CPU and 4.00GB RAM .

Since the convergence of first-order methods may slow down as the iterates approach a stationary point, it is critical to detect this and stop properly. In addition, it is tricky to correctly predict whether an algorithm is temporarily or permanently trapped in a region when its convergence speed has reduced. Hence, it is usually beneficial to have flexible termination rules. In our implementation, in addition to checking the norm of the gradient $\|\mathrm{grad} f(x)\| \le \epsilon_g$, we also compute the relative changes of objective function values of the two consecutive iterates and terminate it as soon as

$$\frac{f(x_k) - f(x_{k+1})}{|f(x_k)| + 1} \le \epsilon_f. \tag{5.1}$$

The default values of ϵ_f, ϵ_g are $10^{-8}, 10^{-5}$. The max iteration is 1000.

Given a symmetric matrix $A \in \mathbb{R}^{n\times n}$, the p-largest eigenvalue problem can be formulated as

$$\begin{aligned} \max_{X\in\mathbb{R}^{n\times p}} \quad & \mathrm{tr}(X^T AX) \\ \text{s.t.} \quad & X^T X = I_p. \end{aligned}$$

We form a few randomly generated dense Wishart matrices assembled as $A = \bar{A}\bar{A}^T$, where $\bar{A} \in \mathbb{R}^{n\times n}$ is a matrix whose elements are sampled from the standard Gaussian distribution. The initial iterate X_0 is given by applying Matlab's function *orth* to a matrix whose elements are drawn from the standard normal distribution using Matlab's function *randn*.

The objective function is constrained on the Stiefel manifold $St(p, n) = \{X \in \mathbb{R}^{n\times p} : X^T X = I_p\}$. The tangent space is $T_X St(p, n) = \{Z \in \mathbb{R}^{n\times p} : X^T Z + Z^T X = 0\}$. We select the gradient, retraction and vector transport as follows. Define the function

$$\bar{f} : \mathbb{R}^{n\times p} \to \mathbb{R} : X \mapsto \mathrm{tr}(X^T AX).$$

Let f denote the restriction of $\bar{f}$ to the Stiefel manifold. We have $D\bar{f}(X)[Z] = 2\mathrm{tr}(Z^T AX)$, hence $\mathrm{grad}\bar{f}(X) = 2AX$. Then the gradient of f is equal to the projection of $\mathrm{grad}\bar{f}(X)$ onto $T_X St(p, n)$:

$$\mathrm{grad} f(X) = P_X \mathrm{grad}\bar{f}(X) = (I - XX^T)\mathrm{grad}\bar{f}(X) + X\mathrm{skew}(X^T \mathrm{grad}\bar{f}(X)),$$

where $\mathrm{skew}(S) := \frac{1}{2}(S - S^T)$. The retraction is

$$R_X(\xi) := qf(X + \xi),$$

where $qf(S)$ denotes the Q factor of the decomposition of S as $S = QR$, where Q belongs to $St(p, n)$ and R is an upper triangular matrix with strictly positive diagonal elements. Since the isometry condition can be dropped on compact manifolds (see [14]), we choose the vector transport as below

$$\mathcal{T}_{X_1,X_2}\xi = (I - X_2X_2^T)\xi + X_2\mathrm{skew}(X_2^T\xi) \in T_{X_2}St(p, n),$$

where $X_2 = R_{X_1}(\eta), \xi, \eta \in T_{X_1}St(p, n)$.

We conduct our numerical experiments on the problem above. For simplicity, we describe the Riemannian subspace quasi-Newton algorithm as RSQN 1 and RSQN 2, where we choose ρ_k by (3.10) and (3.8) respectively. The Riemannian steepest descent method is abbreviated as 'RSD' and the Riemannian Conguate Gradient method is called 'RCG' for short (see [5,8]). We record the average numerical performance and list them in Table 1, 2 in which 'iter' represents the iteration number, 'CPU' represents the required time, 'obj' represents the objective function value, 'nf' represents the number of the function evaluations and 'ng' represents the number of gradient evaluations.

Table 1: Numerical results of RSD and RSQN1

n, p	RSD					RSQN1				
	CPU	iter	obj	nf	ng	CPU	iter	obj	nf	ng
$n = 100$, various p (I)										
$p = 3$	0.25	113	542.0168	367	114	0.07	46	542.0168	192	192
$p = 5$	0.35	132	874.2449	444	133	0.10	50	874.2449	195	195
$p = 10$	0.40	148	1586.8262	484	149	0.12	50	1586.8262	192	192
$p = 5$, various n (II)										
$n = 100$	0.37	145	873.6855	440	146	0.09	43	873.6855	172	172
$n = 500$	3.96	336	4768.8005	961	337	0.42	81	4768.8005	236	236
$n = 1000$	14.72	448	9716.6753	1188	449	1.26	103	9716.6753	219	219
$n = 100, p = 5$, various $cond(A)$ (III)										
$O(10^4)$	0.27	106	864.3154	343	107	0.10	50	864.3154	194	194
$O(10^5)$	0.29	109	864.9386	367	110	0.10	48	864.9386	189	189
$O(10^6)$	0.40	147	882.5408	479	148	0.09	49	882.5408	178	178

Table 2: Numerical results of RCG and RSQN2

n, p	RCG					RSQN2				
	CPU	iter	obj	nf	ng	CPU	iter	obj	nf	ng
$n = 100$, various p (I)										
$p = 3$	0.12	53	542.0168	154	54	0.07	47	542.0168	180	180
$p = 5$	0.16	58	874.2449	168	59	0.08	48	874.2449	170	170
$p = 10$	0.20	64	1586.8262	186	65	0.12	57	1586.8262	212	212
$p = 5$, various n (II)										
$n = 100$	0.17	59	873.6855	172	60	0.08	43	873.6855	151	151
$n = 500$	1.12	93	4768.8005	261	94	0.35	79	4768.8005	176	176
$n = 1000$	3.71	108	9716.6753	301	109	1.17	101	9716.6753	196	196
$n = 100, p = 5$, various $cond(A)$ (III)										
$O(10^4)$	0.16	58	864.3154	167	59	0.10	51	864.3154	195	195
$O(10^5)$	0.16	55	864.9386	164	56	0.09	48	864.9386	168	168
$O(10^6)$	0.17	60	882.5408	172	61	0.09	49	882.5408	178	178

Table 1, 2 contain the results with various n, p and $cond(A)$ for the RSD, RSQN1, RCG, RSQN2 over random tests. For a fixed n, it is clear that from Table 1(I), 2(I), our subspace algorithms, RSQN1 and RSQN2, perform more efficient than RSD, and RCG in terms of CPU time and iterations for small p. Since we adopt the Wolfe line search in our subspace methods, RSQN1 and RSQN2 require more gradient evaluations than RCG. For a fixed p,

Table 1(II), 2(II) show that the advantage of the RSQN1, RSQN2 is more obvious especially when n grows larger in terms of CPU time, iterations and the number of function evaluations, which indicate the efficiency of the subspace method. For $n = 100, p = 5$, we investigate the influence of the condition number of the random matrix A on the algorithm in Table 1(III), 2(III). It is clear that the CPU time, iterations, the number of function evaluations of our algorithm keep stable as the condition numbers grow. Overall, our algorithm is efficient and stable in most cases, even for ill-conditioned problems.

References

[1] P.-A. Absil, C.G. Baker and K.A. Gallivan, A truncated CG style method for symmetric generalized eigenvalue problems, *J. Comput. Appl. Math.* 189 (2006) 274–285.

[2] P.-A. Absil, C.G. Baker and K.A. Gallivan, Trust-region methods on Riemannian manifolds, *Found. Comput. Math.* 7 (2007) 303–330.

[3] R.L. Adler, J.P. Dedieu, J.Y. Margulies, M. Martens and M. Shub, Newton's method on Riemannian manifolds and a geometric model for the human spine, *IMA J. Numer. Anal.* 22 (2002) 359–390.

[4] P.-A. Absil and K.A. Gallivan, Accelerated line-search and trust-region methods, *SIAM J. Numer. Anal.*, 47 (2009) 997–1018.

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

[6] C.G. Baker, *Riemannian manifold trust-region methods with applications to eigenproblems*, PhD thesis, School of Computational Science, Florida State University, 2008.

[7] C.G. Baker, P.-A. Absil and K.A. Gallivan, An implicit trust-region method on Riemannian manifolds, *IMA J. Numer. Anal.* 28 (2008) 665–689.

[8] N. Boumal, B. Mishra, P.-A. Absil and R. Sepulchre, Manopt, a Matlab Toolbox for Optimization on Manifolds, *J. Mach. Learn. Res.* 15 (2014) 1455–1459.

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

[10] S. Hosseini and M.R. Pouryayevali, Generalized gradients and characterization of epi-Lipschitz sets in Riemannian manifolds, *Nonlinear Anal.* 74 (2011) 3884–3895.

[11] Y. Ledyaev and Q. Zhu, Nonsmooth analysis on smooth manifolds, *Trans. Amer. Math. Soc.* 359 (2007) 3687–3732.

[12] J. Nocedal and S.J. Wright, Numerical optimization, Springer Series in *Operations Research and Financial Engineering*, Springer, New York, second edition, 2006.

[13] C. Qi, *Numerical Optimization Methods On Riemannian Manifolds*, PhD thesis, Florida State University, 2011.

[14] W. Ring and B. Wirth, Optimization methods on Riemannian manifolds and their application to shape space, *SIAM J. Optim.* 22 (2012) 596–627.

[15] B. Simon, Trace ideals and their applications, in *Mathematical Surveys and Monographs American Mathematical Society*, Vol. 120, Providence, RI, second edition, 2005.

[16] S.T. Smith, Optimization techniques on Riemannian manifolds, in *Hamiltonian and gradient flows, algorithms and control*, volume 3 of Fields Inst. Commun., Amer. Math. Soc., Providence, RI, 1994, pp. 113–136.

[17] C. Udriste, Kuhn-Tucker theorem on Riemannian manifolds, in *Topics in differential geometry*, North-Holland, Amsterdam, Vol. I, II (Debrecen, 1984), 1988, pp. 1247–1259.

[18] C. Udriste, *Convex Functions and Optimization Methods on Riemannian Manifolds*, Kluwer Academic Publishers Group, Dordrecht, 1994.

[19] Z. Wen and W. Yin, A feasible method for optimization with orthogonality constraints, *Math. Program.* 142 (2013) 397–434.

[20] Y. Yuan and J. Stoer, A subspace study on conjugate gradient algorithms. *Z. Angew. Math. Mech.* 75 (1995) 69–77.

[21] C. Yang, J. Meza and L. Wang, A constrained optimization algorithm for total energy minimization in electronic structure calculations, *J. Comput. Phys.* 217 (2006) 709–721.

[22] W. Yang, L. Zhang and R. Song, Optimality conditions for the nonlinear programming problems on Riemannian manifolds, *Pac. J. Optim.* 10(2) (2014) 415–434.

Manuscript received 23 January 2015
revised 15 March 2016
accepted for publication 21 September 2016

Hejie Wei
School of Mathematical Sciences, Fudan University
Shanghai 200433, China, P.R. China
E-mail address: 12110180023@fudan.edu.cn