Research on Trust Region Algorithms for Orthogonally Constrained Optimization Problems

Xianfeng Ding, Pengfei Wen, Xiaolin Yi, Hanbing Mei, Yiyu Qin and Qianmei Luo

Abstract—In this paper, a nonmonotone adaptive trust region algorithm is proposed to solve a classical optimization problem with orthogonal constraints. Specifically, the optimization problem is transformed into an unconstrained optimization problem on the Stiefel manifold aiming at reducing computational complexity and cost. Theoretical analysis confirms the global and local convergence of this algorithm under specific conditions. Additionally, numerical simulations are performed and the corresponding experimental results demonstrate the effectiveness of our proposed method.

Index Terms—Orthogonal Constraints, Stiefel Manifold, Nonmonotone Adaptive Trust Region Algorithm, Numerical Simulation.

I. Introduction

CONSIDER an optimization problem with orthogonal constraints, which is as follows,

$$\begin{aligned} &\min_{X\in\mathbb{R}^{n\times r}} f(X) \\ &s.t.\ X^TX = I_r \end{aligned} \tag{I.1}$$

where $f(X)$:$\mathbb{R}^{n\times r} \to \mathbb{R}$ is a continuously differentiable real-valued function, $I_r \in \mathbb{R}^{r\times r}$ represents the r-th order identity matrix. The feasible set $St(n,r) := \{X \in \mathbb{R}^{n\times r} : X^TX = I_r\}$, $r \le n$ constrained by orthogonal conditions is referred to as the Stiefel manifold, which is an embedded submanifold in linear space $\mathbb{R}^{n\times r}$ and its dimension is $rn - \frac{1}{2}r(r+1)$. For further details of Stiefel manifold, one can refer to monograph [1], [2], [3], [4]. Specifically, Stiefel manifold will be simplified to an unit sphere S_r^{n-1} when $r = 1$, which dimension is $n = 1$. In this case, problem I.1 is transformed into a spherical constrained optimization problem. In addition, Stiefel manifold can be simplified to an orthogonal group $o(n)$ when $r = n$, which dimension is $\frac{n}{2}(n-1)$.

Manuscript received July 4, 2024; revised February 26, 2025.
This work was supported by the National Social Science Fund of China 22XGL019.

Xianfeng Ding is an associate professor at the School of Science, Southwest Petroleum University, Chengdu, 610500, China (e-mail: dingxianfeng@swpu.edu.cn).
Pengfei Wen is a postgraduate student at the School of Science, Southwest Petroleum University, Chengdu, 610500, China (e-mail: 2522303534@qq.com).
Xiaolin Yi is a postgraduate student at the School of Science, Southwest Petroleum University, Chengdu, 610500, China (e-mail: 2083391532@qq.com).
Hanbing Mei is a postgraduate student at the School of Science, Southwest Petroleum University, Chengdu, 610500, China (e-mail: 1148765513@qq.com).
Yiyu Qin is a postgraduate student at the School of Science, Southwest Petroleum University, Chengdu, 610500, China (e-mail: 184724829@qq.com).
Qianmei Luo is a postgraduate student at the School of Science, Southwest Petroleum University, Chengdu, 610500, China (e-mail: 1570628426@qq.com).

Problem I.1 has widespread applications in various fields including medicine, machine learning, and automation technologies. It encompasses applications such as the Kohn-Sham total energy minimization problem [5], [6], [7], the orthogonal procrustes problem [8], [9], sparse principal component analysis problem [10], [11], eigenvalue problem [12], and matrix singular value decomposition problem [13]. The orthogonal constrained optimization problem exhibits non-convex and non-linear characteristics, which makes analytical computation of it face tremendous challenges. In scientific and engineering applications, the problems are not only large-scale but also involve core mathematical models subject to orthogonality constraints. Currently existing algorithms for solving this problem still have some shortcomings. For instance, these algorithms may only guarantee finding a local optimal solution for most optimization problems, and ensuring that the computation at each iteration point remains within the feasible domain can be computationally expensive, this has sparked a strong research interest among numerous experts and scholars in this field.

The existing algorithms for solving orthogonal constrained optimization problems can be broadly categorized into feasible algorithms [14], [15], [16] and infeasible algorithms [12]. Feasible algorithms require that the iteration point satisfies the orthogonal constraints at each iteration, with a key challenge being the definition of a suitable contraction mapping. Currently, the geodesic-based and projection-based methods are two prominent categories of algorithms for effectively computing the contraction mapping. Geodesic-based methods select an appropriate iteration step size and search for the next iteration point along the geodesic direction of the manifold. In contrast, projection algorithms are a type of algorithm that search for the next iteration point of the current point in the tangent space and project the iteration point back into the manifold. In practical applications of large-scale orthogonal constrained optimization problems, feasible algorithms suffer from a significant increase in computational complexity. In such cases, it is prudent to consider employing appropriate infeasible algorithms to circumvent this drawback of excessive computational demands. Infeasible algorithms do not strictly demand that the points obtained during the iteration process to satisfy the orthogonal constraints. However, the sequence of iterates should gradually converge to a stationary point that satisfies the orthogonal constraints.

II. Preliminary Preparation

In this section, some relevant background on Trust Region algorithm and Stiefel manifold has been intro-

duced, including basic definition and conditions. For more details, one can refer to the monograph [17]. In subsection 2.1, the needed theoretical foundations of Trust Region algorithm will be presented. Specifically, since Stiefel manifold is a typical Riemannian manifold in smooth tangent space, the unconstrained optimization problem on Stiefel manifold can be solved by the existing optimization algorithms on Riemannian manifold, including Riemannian Newton (RN), Riemannian Trust Region (RTR), and Riemannian Stochastic Descent (RSD) algorithms. Different from Trust Region algorithm in Euclidean space, Riemannian trust region algorithm should be provided the manifold information, tangent space and the Riemannian metric, at each iteration. Then, the above information is utilized to calculate the Hessian matrix and step size for each trust region subproblem, ensuring each iteration remains on the manifold. Thus, in subsection 2.2, the needed theoretical foundations of Riemannian manifold optimization will be provided. In Subsection 2.3, the built method will be proposed, that is, the improved nonmonotone adaptive trust region algorithm on manifolds.

A. Trust Region Algorithm

Its basic idea is to transform optimization problems into several local optimization subproblems. In each iteration, a specific trust region is delineated, and the optimal solution is determined within the confines of this trust region of the model. This method solves the minimization problem of a quadratic model in each iteration step and checks whether the function value has improved after updating to adjust the radius of the local region.

The Trust Region algorithm is a numerical method used to solve nonlinear optimization problems, which iteratively captures the minimum value of the objective function. The details of Trust Region algorithm will be presented as follows. By given an unconstrained optimization problem $\min_{x\in\mathbb{R}^n} f(x)$, where $f:\mathbb{R}^n\to\mathbb{R}$ denote the objective function, which is real-valued and quadratic continuously differentiable.

To obtain each iteration step, trust region methods compute a trial step d_k by solving a subproblem:

$$\begin{aligned}&\min_{d\in\mathbb{R}^n} m_k(d)=f(x_k)+g(gradf(x_k),d)+\frac{1}{2}g(H_k[d],d)\\&s.t.g(d,d)\le\Delta_k^2\end{aligned}\tag{II.1}$$

where $g(\cdot,\cdot)$ represents the inner product on $\mathbb{R}^n$, $H_k\in\mathbb{R}^{n\times n}$ is a symmetric approximation of $Hessf(x_k)$, and $\Delta_k>0$ is the trust region radius. Special assumption $g_k=gradf(x_k)$, when $H_k\succ 0$ and $\left\|H_k^{-1}g_k\right\|\le\Delta_k$, the trial step $d_k^H=-H_k^{-1}g_k$ in equation II.1. The ratio ρ_k is defined as:

$$\rho_{\dot{k}}=\frac{Ared_{\dot{k}}}{Pred_{\dot{k}}}=\frac{f(x_{\dot{k}})-f(x_{\dot{k}}+d_{\dot{k}})}{m_{\dot{k}}(0)-m_{\dot{k}}(d_{\dot{k}})}\tag{II.2}$$

in equation II.2, $Ared_k$ is referred to as the actual reduction of the objective function, and $Pred_k$ is the predicted reduction. If ρ_k is close to 1, it indicates that the second-order approximation model closely matches the objective function. In this case, the current iteration step is accepted. If ρ_k is close to 0, it indicates that the second-order approximation model deviates significantly from the actual objective function. The current iteration step is rejected as a result. In the next iteration, the trust region radius should be reduced, and a new trial step d_k needs to be recalculated[18].

B. Standard Results in Riemannian Geometry

Riemannian manifold is mainly based on Riemannian Geometry, thus some basic definitions and foundations of Riemannian geometry will be provided in following [19], [20], [21], [22], [23].

Setting M is a smooth differentiable manifold C^∞, the smooth bijection φ in Euclidean Space is defined as $\forall x\in M,\exists U\subset M,x\in U$, and $\exists V\subset\mathbb{R}^n$, such that $\varphi:U\mapsto V$, where is an open neighborhood containing the point x, V is an open set in the $n-$dimensional Euclidean space $\mathbb{R}^n$.

The tangent vector at point x on Riemannian manifold is formed by the velocity vector of a smooth curve at point x. When the all tangent vectors at point x is denoted as the tangent space T_xM, the tangent bundle of M is constructed by the disjoint union of all tangent spaces, shown as: $TM=\left\{(x,v):x\in M,v\in T_xM\right\}$, where the tangent bundle TM is a vector bundle associated with the Riemannian manifold M. In addition, when a smooth projection is mapped by $\pi:TM\mapsto M$, the base of vector can be extracted by $\pi(x,\nu)=x$.

The Riemannian metric defines an inner product on the tangent space T_xM, which we denote as $g_x(\cdot,\cdot)$:$T_xM\times T_xM\to\mathbb{R}$, for $\forall s,w,u\in T_xM$ and $a,b\in\mathbb{R}$, the inner product $g_x(s,w)$ satisfies symmetry, bilinearity, and positive definiteness, i.e.:

$$\begin{aligned}&(1)g_x(s,w)=g_x(w,s),\\&(2)g_x(as+bw,u)=ag_x(s,u)+bg_x(w,u),\\&(3)g_x(s,s)\ge 0,g_x(s,s)=0\Longleftrightarrow s=0.\end{aligned}$$

Specifically, the Riemannian metric can be represented by defining a symmetric positive definite quadratic form on each tangent space T_xM, the norm of a tangent vector $w\in T_xM$ is denoted as $\|w\|_x=\sqrt{g_x(w,w)}$. Based on [19], an affine connection is defined to describe the way in which tangent spaces on the manifold are connected.

Definition II.1. An affine connection ∇ on a manifold M is a mapping $\nabla:\varepsilon(M)\times\varepsilon(M)\to\varepsilon(M):(X,Y)\to\nabla_XY$, that satisfies the following three properties:

$$\begin{aligned}&(1)\sigma(M)\text{-linear:}\nabla_{fX+gY}Z=f\nabla_XZ+g\nabla_YZ,\\&(2)\mathbb{R}\text{linear:}\nabla_X(aY+bZ)=a\nabla_XY+b\nabla_XZ,\\&(3)\text{Leibniz's Rule:}\nabla_X(fY)=(Xf)Y+f\nabla_XY.\end{aligned}$$

Where $\varepsilon(M)$ represents the set of smooth vector fields on the manifold M, $\sigma(M)$ denotes the set of smooth scalar fields on the manifold, the vector field ∇_XY represents the covariant derivative of Y with respect to x and is associated with the corresponding affine connection ∇, $X,Y,Z\in\varepsilon(M)$, $f,g\in\sigma(M)$ and $a,b\in\mathbb{R}$, the affine connection ∇ is also known as the Levi-Civita connection or Riemannian connection.

A smooth curve defined on manifold M, where $p, \tilde{p} \in \mathbb{R}$. The length of this curve is denoted as $L(\gamma) = \int_p^{\tilde{p}} \|\gamma'(t)\|_{\gamma(t)} dt$, and is used to connect points x and y on the manifold M, i.e., $\gamma(p) = x$ and $\gamma(\tilde{p}) = y$. The curve $\gamma(t) = x$ is a geodesic on M if and only if $\nabla_{\gamma'(t)}\gamma'(t) = 0$ holds, $\nabla_{\gamma'(t)}\xi(t) = 0$ represents the parallel transport of the tangent vector $\xi(t)$ along the curve $\gamma(t)$, where ∇ is the Levi-Civita connection on the manifold, $\gamma'(t)$ is the tangent vector of the curve γ, and $\nabla_{\gamma'(t)}\gamma'(t)$ represents the covariant derivative of the tangent vector.

Definition II.2. Riemannian distance on manifold M:

$$d_R : M \times M \to \mathbb{R}^+ : (x, y) \to d_R(x, y) = \inf_{\gamma \in \varsigma} L(\gamma),$$

where ς is the set of all C^1 curves $\gamma : [p, \tilde{p}] \to M$.

Definition II.3. Let f be a scalar function on manifold M. The gradient of f at point $x \in M$, denoted as $gradf(x)$, is defined as the unique element in T_xM satisfying:

$$g_x\Big(gradf\Big(x\Big), \xi\Big) = Df\Big(x\Big)\Big[\xi\Big], \forall \xi \in T_xM.$$

Where $Df(x)[\xi]$ represents the directional derivative of f at point x in the direction of ξ.

Definition II.4. For a given scalar function f on manifold M, the Riemannian Hessian operator of f at point $x \in M$, is a linear mapping $Hessf(x)$ from T_xM to itself, satisfying:

$$\nabla_\xi gradf(x) = Hessf(x)[\xi], \forall \xi \in T_xM,$$

where ∇ represents the Levi-Civita connection. In fact, the Riemannian Hessian operator is a symmetric operator with respect to the Riemannian metric, i.e.:

$$g_x(Hessf(x)[\xi], \eta) = g_x(\xi, Hessf(x)[\eta]), \forall \xi, \eta \in T_xM.$$

One of the important applications of the exponential map in Riemannian geometry is to map tangent vectors in the tangent space to curves or points on the manifold, and project them along geodesics on the manifold. Specifically, for $\forall x \in M$ and an arbitrary tangent vector $\varpi \in T_xM$ at point x, the exponential map projects ϖ onto a point on the manifold M which is obtained by moving along the direction of the tangent vector ϖ from the point x by a specified distance. The exponential map at point x is defined as follows:

$$Exp_x : T_xM \to M : \xi \to Exp_x\xi = \gamma(1; x, \xi),$$

where $\xi \in T_xM$, and γ is a geodesic that satisfies $\gamma(0) = x$ and $\gamma'(0) = \xi$.The application of the exponential map allows us to uniquely define local trust region subproblems on the manifold by locally mapping the Riemannian manifold to the Euclidean space T_xM [24].

The prerequisite for the exponential map is the requirement to calculate geodesics, which usually involves solving differential equation problems. This leads to an escalation in computational cost [25]. In practical applications, systematic use of the exponential map is not always applicable in every situation. Therefore, we substitute the exponential map with a class of mappings known as retractions [1], [26], [27], [28], [29]. On one hand, retractions no longer rely on the curve γ being a geodesic, which significantly reduces computational costs. On the other hand, retractions exhibit most of the properties of the exponential map in the optimization process.

Definition II.5. The retraction R on the manifold M is a C^2 smooth mapping from the tangent bundle TM to M. Given $x \in M$, let $R_x : T_xM \to M$ denote the restriction of R to T_xM, satisfying:

(1)R is continuously differentiable,

(2)$R_x(0_x) = x$, where0_xdenotes the zero element ofT_xM,

(3)$DR_x(0_x) = id_{T_xM}$,

where id denotes the identity mapping on T_xM, possessing the canonical identification $T_{0_x}T_xM \simeq T_xM$.

In a manifold, parallel transport can move a tangent vector from one tangent space to another while preserving the original information of the tangent vector. This allows for the comparison of tangent vectors from different points by transporting them to a common tangent space. Similar to retractions serving as a substitute for the exponential map, parallel transport incurs high computational costs. Hence, we contemplate utilizing retraction-based vector transport as an alternative to parallel transport, as referenced in [1].

Definition II.6. The vector transport associated with the retraction R_x is a smooth mapping:

$$V : TM \oplus TM \to TM : (\eta_x, \xi_x) \mapsto T_{\eta_x}(\xi_x),$$

where $TM \oplus TM = \{(\eta_x, \xi_x) : \eta_x, \xi_x \in T_xM, x \in M\}$, and satisfies the properties of contraction adjoint ness, consistency, and linearity, i.e.:

$$\begin{aligned}
&(1)V_{\eta_x}(\xi_x) \in T_{R_x(\eta_x)}M,\\
&(2)V_{0_x}(\xi_x) = \xi_x, \quad \forall \xi_x \in T_xM,\\
&(3)V_{\eta_x}(a\xi_x + b\zeta_x) = aV_{\eta_x}(\xi_x) + bV_{\eta_x}(\zeta_x),
\end{aligned}$$

where $\forall \eta_x, \xi_x, \zeta_x \in T_xM$ and $a, b \in \mathbb{R}$.

In particular, if the vector transport V also satisfies

$$g_x\left(V_{E(\eta_x)}(\xi_x), V_{E(\eta_x)}(\zeta_x)\right) = g_x(\xi_x, \zeta_x),$$

it is referred to as an isometric vector transport, denoted as V_E. Furthermore, we use V_{R_x} to denote the differential retraction, i.e.:

$$V_{R_x(\eta_x)}\xi_x = DR_x(\eta_x)[\xi_x] = \frac{d}{dt}R_x(\eta_x + t\xi_x)\Big|_{t=0}.$$

C. Improved Nonmonotone Adaptive Trust Region Algorithm on Manifolds

If each obtained point in any iteration is feasible, the problem I.1 is a typical unconstrained optimization problem on the Stiefel manifold. Specifically, since any orthogonal matrix can be regarded as a point on the Stiefel manifold, i.e., $Q \in St(n, r) := \{X \in \mathbb{R}^{n\times r} : X^TX = I_r\}$, the problem I.1 can be transformed into as an unconstrained optimization problem on the manifold, shown as

$$\begin{aligned}
&\min_{Q \in St(n,r)} f(Q) = \min_{X \in \mathbb{R}^{n\times r}} f(X)\\
&s.t.\ X^TX = I_r
\end{aligned} \tag{II.3}$$

In the Riemannian Trust-Region (RTR) algorithm, the approximated model $\hat{m}_{x_k} := m_{x_k} \circ R_{x_k}$ around $x_k \in M$ of objective function $\hat{f}_{x_k}$ in problem II.1 is obtained by the second-order Taylor expansion of $\hat{f}_{x_k} := f_{x_k} \circ R_{x_k}$ [30]. Subsequently, by using the retraction R_{x_k}, the minimization problem of f_{x_k} on manifold M is locally mapped to the minimization problem of the objective function:

$$\hat{f}_{x_k} : T_{x_k}M \to R : \xi \mapsto f_{x_k}\left(R_{x_k}\left(\xi\right)\right),$$

where $\hat{f}_{x_k}$ is a real-valued function on $T_{x_k}M$.Meanwhile, the trust region subproblem on $T_{x_k}M$ is defined as,

$$\begin{aligned} &\min_{d \in T_{x_k}M} \hat{m}_{x_k}(d) = \hat{f}_{x_k}(x_k) + g_x\left(grad\hat{f}_{x_k}(x_k), d\right) \\ &+ \frac{1}{2} g_x\left(H_{x_k}[d], d\right) \\ &s.t.\ g_x(d,d) \le \hat{\Delta}_k^2 \end{aligned} \tag{II.4}$$

where $H_{x_k} : T_{x_k}M \to T_{x_k}M$ is a symmetric linear operator, i.e.,

$$g_x\left(H_{x_k}\xi, \chi\right) = g_x\left(\chi, H_{x_k}\xi\right), \forall \xi, \chi \in T_xM. \tag{II.5}$$

To simplify the notation, $\hat{m}(d)$, $\hat{f}(x_k)$ and $\hat{g}_k$ are applied to represent $\hat{m}_{x_k}(d), \hat{f}_{x_k}(x_k)$ and $grad\hat{f}_{x_k}(x_k)$ respectively in the context. Let q_k satisfy

$$q_k = \begin{cases} -\hat{g}_k, & if k = 0 or \frac{-g_x(\hat{g}_k, d_{k-1})}{\|\hat{g}_k\| \cdot \|d_{k-1}\|} \le \tau \\ d_{k-1}, & otherwise \end{cases}, \tag{II.6}$$

where d_{k-1} is the solution of the trust region subproblem II.4, $\tau \in (0,1)$ is a constant. To prevent the radius of trust region from being too small, s_k is supposed as

$$s_k = \begin{cases} -\frac{g_x(\hat{g}_k, q_k)}{q_k^T H_{x_k} q_k}\|q_k\|, & if k = 0 \\ \max\left\{-\frac{g_x(\hat{g}_k, q_k)}{q_k^T H_{x_k} q_k}\|q_k\|, \lambda\hat{\Delta}_{k-1}\right\}, & if k \ge 1 \end{cases}, \tag{II.7}$$

where $\lambda > 1$. Thus, the radius of trust region is shown as

$$\overline{\Delta}_{\dot{\kappa}} = \rho^{\alpha} \min\left\{s_{\dot{\kappa}}, \kappa\right\}, \tag{II.8}$$

where $\rho \in (0,1)$, α is a non-negative integer, and $\kappa > 0$ is a real-valued constant.

Therefore, we need to solve the following subproblem in the nonmonotone trust region algorithm at the iteration point x_k.

$$\begin{aligned} &\min_{d \in T_{x_k}M} \hat{m}(d) = \hat{f}(x_k) + g_x(\hat{g}_k, d) + \frac{1}{2} g_x\left(H_{x_k}[d], d\right) \\ &s.t.\ g_x(d,d) \le \left(\rho^{\alpha} \min\left\{s_{\dot{\kappa}}, \kappa\right\}\right)^2 \end{aligned} \tag{II.9}$$

Similar to trust region algorithms in Euclidean spaces, the trial step $\hat{d}_k$ obtained through subproblem II.9 holds the following definition,

$$\hat{\rho}_{\dot{k}} = \frac{D_{\dot{k}} - \hat{f}\left(R_{x_k}\left(\hat{d}_{\dot{k}}\right)\right)}{\hat{m}(0) - \hat{m}\left(\hat{d}_{\dot{k}}\right)} = \frac{D_{\dot{k}} - \hat{f}\left(x_{\dot{k}+1}\right)}{\hat{m}(0) - \hat{m}\left(\hat{d}_{\dot{k}}\right)}. \tag{II.10}$$

The purpose of approximating $\hat{f}(x_k) - \hat{f}\left(R_{x_k}\left(\hat{d}_k\right)\right)$ by $D_k - \hat{f}\left(R_{x_k}\left(\hat{d}_k\right)\right)$ in equation II.10 is to prevent the objective function from monotonically decreasing during the iteration process, while ensuring that D_k satisfies

$$D_{\dot{k}} = \begin{cases} \hat{f}(x_{\dot{k}}), & if k = 0 \\ \eta_{\dot{k}} D_{k-1} + (1 - \eta_{\dot{k}})\hat{f}(x_{\dot{k}}), & if k \ge 1 \end{cases}, \tag{II.11}$$

where $\eta_k \in [\eta_{\min}, \eta_{\max}], \eta_{\min} \in (0,1)$, and $\eta_{\max} \in [\eta_{\min}, 1)$. Consider $\hat{\rho}_{\dot{\kappa}} \ge \mu$ and $\mu \in (0,1)$, and set $x_{k+1} = x_{k+1}^{+} = R_{x_k}\left(\hat{d}_{\dot{\kappa}}\right)$, the iteration of RTR is ensured feasibility [31], [32], [33], [34], [35], [36], [37].

The detail of nonmonotone adaptive trust region algorithm on manifolds is shown in Algorithm 1.

In practical applications, the dimension l of the manifold M is very large, which makes solving a linear system $(H_{x_k} + \mu^* id)\, d_{k-1} = -\hat{g}_k$ of size l or check the positive definiteness of an l×l matrix $H_{x_k} + \mu^* id$ is infeasible, where d_{k-1} is the global solution to problem II.4 if and only if $\mu^* > 0$ [1]. The existed algorithms aiming at alleviating numerical burdens, the truncated conjugate gradient method (t-CG) is one of the most popular [24], which is designed to ensure that when $H_{x_k} \succ 0$ and $\left\|H_{x_k}^{-1}\hat{g}_k\right\| \le \Delta_k$, we can obtain an approximate solution $d_{k-1} = -\left(H_{x_k}\right)^{-1}\hat{g}_k$ to subproblem II.4 under the sufficient number of iterations. In addition, for the proposed algorithm, it is necessary to improve the similarity compared with the Cauchy point in each iteration, which can ensure that the algorithm gradually obtain the optimal solution. Thus, the inequation $\left\|\hat{d}_k\right\| \le \rho^{\alpha} \min\left\{s_k, \kappa\right\}$ will be satisfied, i.e.,

$$\hat{m}(0) - \hat{m}\left(\hat{d}_k\right) \ge \beta \|\hat{g}_k\| \min\left\{\overline{\Delta}_k, \frac{\|\hat{g}_k\|}{\|H_{x_k}\|}\right\}, \tag{II.12}$$

$$g_x\left(\hat{d}_k, \hat{g}_k\right) \le -\beta\|\hat{g}_k\| \min\left\{\overline{\Delta}_k, \frac{\|\hat{g}_k\|}{\|H_{x_k}\|}\right\}, \tag{II.13}$$

where $\beta \in (0,1)$.

III. The Convergence

This section provides an overview w.r.t. the convergence of proposed algorithm, including global convergence and local convergence.

A. Global Convergence

Since the objective function is quadradic continuously differential and bounded in manifold space, we provide the following standard assumptions to make the global convergence identifiable.

Assumption 1. Given a level set $C(x_0) = \{x \in M : f(x) \le f(x_0)\} \subset \Omega$, $x_0 \in M$, where Ω is an open convex set. When the objective function f is quadradic continuously differentiable on $C(x_0)$ and has a lower bounded value on M, $gradf(x)$ is uniformly continuous on Ω.

Assumption 2. For all $x_k \in \mathbb{N} \cup \{0\}$, there exists a positive constant Q_1 such that the symmetric linear operator H_{x_k} is uniformly bounded, i.e., $\|H_{x_k}\| \le Q_1$.

Algorithm 1 Nonmonotone Adaptive Trust Region Algorithm on Manifolds(NATRAM)

Input: Riemannian manifold (M, g), retraction R. Set $x_0 \in M, \mu, \rho, \tau \in (0,1)$, $\lambda > 1, \kappa > 0, \hat{\Delta}_k > 0, \delta > 0, D_0 = \hat{f}(x_0), \eta_k \in [\eta_{\min}, \eta_{\max}]$ with $\eta_{\min} \in (0,1)$ and $\eta_{\max} \in [\eta_{\min}, 1), k := 0, \alpha := 0$

Step 1. If $\|\hat{g}_k\| \leq \delta$, then stop, otherwise, go to step2,

Step 2. Solve Eq.II.4 to determine d_{k-1}, compute q_k according to Eq.II.6, and calculate q_k through Eq.II.7, go to step 3,

Step 3. Compute the trust region radius $\bar{\Delta}_k$ through Eq.II.8, solve Eq.II.9 to determine $\hat{d}_k$, and set $x_{k+1}^+ = R_{x_k}\left(\hat{d}_k\right)$, go to step 4,

Step 4. Compute $\hat{\rho}_k$ using Eq.II.10. If $\hat{\rho}_{\dot{\kappa}} < \mu$, set $\alpha := \alpha + 1$ and go to step 3, otherwise, proceed to step 5,

Step 5. Set $x_{k+1} = x_{k+1}^+$ and compute D_{k+1} using Eq.II.11. Update H_{x_k} using the BFGS quasi-Newton method according to [27]. Set $k := k + 1$ and go back to step 1.

Establishing the global convergence of Algorithm 1 demonstrates that, under appropriate assumptions, the sequence $\{x_k\}_{k\geq 0}$ generated by Algorithm 1 satisfies $\lim_{k\to\infty}\|\hat{g}_k\| = 0$. The following crucial conclusion is a necessary result in proving the global convergence of Algorithm 1.

Our approach to global convergence essentially consists of the nonlinear optimization in Trust Region algorithm. This approach is justified by the following lemma.

Lemma III.1. [24] Suppose that the injectivity radius of the Riemannian manifold (M, g) is $i(M) > 0$, and the real-valued function f on M is Lipschitz continuously differentiable. If $\forall x, y \in M$, holding that $d_R(x, y) < i(M)$, then

$$\left\|P_\gamma^{0\leftarrow 1} gradf(y) - gradf(x)\right\| \leq \beta_1 d_R(y, x), \quad \text{(III.1)}$$

where γ is the unique geodesic with $\gamma(0) = x$ and $\gamma(1) = y$, the vector $P_\gamma^{0\leftarrow 1} gradf(y)$ in $T_x M$ can be obtained by parallel transporting $gradf(y)$ along γ, $\beta_1 > 0$ is a constant, and $d_R(\cdot,\cdot)$ represents the Riemannian distance.

Lemma III.2. Suppose that d_{k-1} is the solution of Trust Region subproblem II.4, and the trust region radius $\bar{\Delta}_k$ is given by II.8. According to Assumption 2, for all $k \in \mathbb{N}$ we have

$$\hat{m}(0) - \hat{m}\left(\hat{d}_{\dot{k}}\right) \geq \frac{1}{2}\rho^\alpha \min\left\{\frac{1}{Q_1}\left(\frac{-g_x\left(\hat{g}_{\dot{k}}, q_{\dot{k}}\right)}{\left\|q_{\dot{k}}\right\|}\right)^2, \kappa\left(\frac{-g_x\left(\hat{g}_{\dot{k}}, q_{\dot{k}}\right)}{\left\|q_{\dot{k}}\right\|}\right)\right\} \quad \text{(III.2)}$$

Proof: Due to $\overline{\Delta}_{\dot{\kappa}} = \rho^\alpha \min\{s_{\dot{\kappa}}, \kappa\}$, the proof of Lemma III.2 mainly consists of two cases,

Case 1: If $s_k \leq \kappa$, then $\overline{\Delta}_{\dot{\kappa}} = \rho^\alpha s_{\dot{\kappa}}$. By given the value $s_{\dot{\kappa}} (k \geq 1)$ shown in equation II.7, we have

$$\begin{aligned}\overline{\Delta}_k &= \rho^\alpha \max\left\{-\frac{g_x(\hat{g}_k, q_k)}{q_k^T H_{x_k} q_k}\|q_k\|, \lambda\hat{\Delta}_{k-1}\right\} \\ &\geq -\rho^\alpha \frac{g_x(\hat{g}_{\dot{\kappa}}, q_{\dot{\kappa}})}{q_{\dot{\kappa}}^T H_{x_k} q_{\dot{\kappa}}}\|q_{\dot{\kappa}}\|.\end{aligned} \quad \text{(III.3)}$$

Then, it implies that

$$d_{\dot{\kappa}}^* = -\rho^\alpha \frac{g_x(\hat{g}_{\dot{\kappa}}, q_{\dot{\kappa}})}{q_{\dot{\kappa}}^T H_{x_k} q_{\dot{\kappa}}} q_{\dot{\kappa}} \quad \text{(III.4)}$$

is a feasible solution of Trust Region subproblem II.9.

According to the Assumption 2, we have $\|H_{x_k}\| \leq Q_1$, thus we obtain that

$$\begin{aligned}&\hat{m}(0) - \hat{m}\left(\hat{d}_{\dot{k}}\right) \geq \hat{m}(0) - \hat{m}\left(d_{\dot{k}}^*\right) \\ &= -g_x\left(\hat{g}_{\dot{k}}, d_{\dot{k}}^*\right) - \frac{1}{2} g_x\left(H_{x_k}\left[d_{\dot{k}}^*\right], d_{\dot{k}}^*\right) \\ &= \rho^\alpha \frac{\left(g_x\left(\hat{g}_k, q_k\right)\right)^2}{q_k^T H_{x_k} q_k}\left(1 - \frac{1}{2}\rho^\alpha\right) \\ &\geq \rho^\alpha \frac{\left(g_x\left(\hat{g}_k, q_k\right)\right)^2}{q_k^T H_{x_k} q_k}\left(1 - \frac{1}{2}\right) \\ &= \frac{1}{2}\rho^\alpha \frac{\left(g_x\left(\hat{g}_{\dot{k}}, q_{\dot{k}}\right)\right)^2}{q_{\dot{k}}^T H_{x_{\dot{k}}} q_{\dot{k}}} \geq \frac{\rho^\alpha}{2Q_1}\left(\frac{g_x\left(\hat{g}_{\dot{k}}, q_{\dot{k}}\right)}{\left\|q_{\dot{k}}\right\|}\right)^2.\end{aligned} \quad \text{(III.5)}$$

Case 2: If $s_k > \kappa$, then $\overline{\Delta}_{\dot{\kappa}} = \rho^\alpha \kappa$. Considering $-\frac{g_x\left(\hat{g}_k, q_k\right)}{q_k^T H_{x_k} q_k}\|q_k\| \leq \kappa$ shown in equation II.7, we have

$$d_{\dot{\kappa}}^* = -\rho^\alpha \frac{g_x(\hat{g}_{\dot{\kappa}}, q_{\dot{\kappa}})}{q_{\dot{\kappa}}^T H_{x_k} q_{\dot{\kappa}}} q_{\dot{\kappa}} \quad \text{(III.6)}$$

where d_k^* is a feasible solution of Trust Region subproblem II.9.

From the equation III.5, it holds that

$$\hat{m}(0) - \hat{m}\left(\hat{d}_{\dot{k}}\right) \geq \frac{\rho^\alpha}{2Q_1}\left(\frac{g_x\left(\hat{g}_{\dot{k}}, q_{\dot{k}}\right)}{\left\|q_{\dot{k}}\right\|}\right)^2. \quad \text{(III.7)}$$

Accordingly, considering $-\frac{g_x\left(\hat{g}_k, q_k\right)}{q_k^T H_{x_k} q_k}\|q_k\| > \kappa$ shown in equation II.7, we can have that

$$-\frac{q_k^T H_{x_k} q_k \kappa}{g_x(\hat{g}_k, q_k)\|q_k\|} < 1. \quad \text{(III.8)}$$

Hence, the point $d_{\dot{\kappa}}^+ = \rho^\alpha \frac{q_{\dot{\kappa}}\kappa}{\|q_{\dot{\kappa}}\|}$ is a feasible solution of Trust Region subproblem II.9. Suppose the equation III.8 holds, we have

$$\begin{aligned}&\hat{m}(0) - \hat{m}\left(\hat{d}_k\right) \geq \hat{m}(0) - \hat{m}\left(d_k^+\right) \\ &= -g_x\left(\hat{g}_k, d_k^+\right) - \frac{1}{2} g_x\left(H_{x_k}\left[d_k^+\right], d_k^+\right) \\ &\geq \rho^\alpha \kappa\left(\frac{-g_x\left(\hat{g}_k, q_k\right)}{\left\|q_k\right\|}\right)\left(1 - \frac{1}{2}\rho^\alpha\right) \\ &\geq \frac{1}{2}\rho^\alpha \kappa\left(\frac{-g_x\left(\hat{g}_k, q_k\right)}{\left\|q_k\right\|}\right).\end{aligned} \quad \text{(III.9)}$$

Thus, the conclusion is established. ■

Lemma III.3. Suppose that the sequence $\{x_k\}_{k\geq 0}$ is generated by Algorithm 1, then we have

$$\forall k \in \mathbb{N}, \hat{f}\left(R_{x_k}\left(\hat{d}_k\right)\right) \leq D_{k+1} \leq D_k \tag{III.10}$$

Proof: Assume that $F = \left\{k : \hat{\rho}_k \geq \mu\right\}$ and $G = \left\{k : \hat{\rho}_k < \mu\right\}$, and suppose D_k in equation II.11 holds, we can obtain

$$\begin{aligned} D_{k+1} - D_k &= \left(1-\eta_{k+1}\right)\hat{f}\left(R_{x_k}\left(\hat{d}_k\right)\right) + \eta_{k+1}D_k - D_k \\ &= \left(\eta_{k+1}-1\right)\left(D_k - \hat{f}\left(R_{x_k}\left(\hat{d}_k\right)\right)\right) \end{aligned} \tag{III.11}$$

and

$$D_{k+1} - \hat{f}\left(R_{x_k}\left(\hat{d}_k\right)\right) = \eta_{k+1}\left(D_k - \hat{f}\left(R_{x_k}\left(\hat{d}_k\right)\right)\right). \tag{III.12}$$

In addition, the proof of Lemma III.3 mainly consists of two cases,

Case 1: Suppose that $k \in F$, equations II.10 and III.2 holds, we have

$$\begin{aligned} &D_k - \hat{f}\left(R_{x_k}\left(\hat{d}_k\right)\right) \\ &= D_{\dot{k}} - \hat{f}\left(x_{\dot{k}+1}\right) \geq \mu\left(\hat{m}\left(0\right) - \hat{m}\left(\hat{d}_k\right)\right) \\ &\geq \frac{1}{2}\mu\rho^{\alpha}\min\left\{\frac{1}{Q_1}\left(\frac{-g_x\left(\hat{g}_{\dot{k}}, q_{\dot{k}}\right)}{\left\|q_{\dot{k}}\right\|}\right)^2, \kappa\left(\frac{-g_x\left(\hat{g}_{\dot{k}}, q_{\dot{k}}\right)}{\left\|q_{\dot{k}}\right\|}\right)\right\} \\ &\geq 0 \end{aligned} \tag{III.13}$$

Accordingly, by combining III.11, III.12, and III.13, we can obtain that

$$\hat{f}\left(R_{x_k}\left(\hat{d}_k\right)\right) \leq D_{k+1} \leq D_k, \forall k \in F. \tag{III.14}$$

Case 2: Suppose that $k \in G$ and $k-1 \in F$. When the equation III.14 holds, we obtain $\hat{f}\left(x_{(k-1)+1}\right) \leq D_k$. Then $\hat{f}\left(R_{x_k}\left(\hat{d}_k\right)\right) = \hat{f}\left(x_k\right)$ is satisfied, then

$$\begin{aligned} D_{k+1} &= \left(1-\eta_{k+1}\right)\hat{f}\left(R_{x_k}\left(\hat{d}_k\right)\right) + \eta_{k+1}D_k \\ &\geq \left(1-\eta_{k+1}\right)\hat{f}\left(R_{x_k}\left(\hat{d}_k\right)\right) + \eta_{k+1}\hat{f}\left(R_{x_k}\left(\hat{d}_k\right)\right) \\ &= \hat{f}\left(R_{x_k}\left(\hat{d}_k\right)\right). \end{aligned} \tag{III.15}$$

Combining III.11, III.12, with III.15, we have

$$\hat{f}\left(R_{x_k}\left(\hat{d}_k\right)\right) \leq D_{k+1} \leq D_k, \forall k-1 \in F. \tag{III.16}$$

If $k-1 \in G$, set $\mathfrak{J} = \left\{u : 1 < u < k, k-u \in F\right\}$. When $\mathfrak{J} = \emptyset$, the function satisfies $\hat{f}\left(R_{x_k}\left(\hat{d}_k\right)\right) = \hat{f}\left(x_{k-i}\right) = \hat{f}\left(x_0\right)$, where $i = 0, 1, 2, \cdots, k-1$. Therefore, according to the D_k shown in equation II.11, $\hat{f}\left(R_{x_k}\left(\hat{d}_k\right)\right) = D_{k+1} = D_k$. When $\mathfrak{J} \neq \emptyset$, set $b = \min\{u : u \in \mathfrak{J}\}$, the function satisfies $\hat{f}\left(R_{x_k}\left(\hat{d}_k\right)\right) = \hat{f}\left(x_{k-l}\right) = \hat{f}\left(x_k\right)$, where $l = 0, 1, 2, \cdots, u-1$, and $k-b \in F$. Therefore, by employing III.14, we obtain $\hat{f}(x_{k-b+1}) \leq D_{k-b+1} \leq D_{k-b}$.

Simultaneously, when

$$\begin{aligned} D_{k-b+2} &= \left(1-\eta_{k-b+2}\right)\hat{f}\left(x_{k-b+2}\right) + \eta_{k-b+2}D_{k-b+1} \\ &\geq \left(1-\eta_{k-b+2}\right)\hat{f}\left(x_{k-b+2}\right) + \eta_{k-b+2}\hat{f}\left(x_{k-b+2}\right) \\ &= \hat{f}\left(x_{k-b+2}\right). \end{aligned} \tag{III.17}$$

Holds, by utilizing III.11, III.12, and III.17, we can deduce that $\hat{f}\left(R_{x_k}\left(\hat{d}_k\right)\right) \leq D_{k+1} \leq D_k$. Combining Cases 1 and 2, the conclusion is established. ■

Lemma III.4. Suppose that the sequence $\{x_k\}_{k\geq 0}$ is generated by Algorithm 1. Then, we have

$$\left|\hat{f}\left(R_{x_k}\left(\hat{d}_k\right)\right) - \hat{f}\left(x_k\right) - \hat{m}\left(0\right) + \hat{m}\left(\hat{d}_k\right)\right| \leq o\left(\left\|\hat{d}_k\right\|^2\right). \tag{III.18}$$

Proof: The proof has been shown in reference [38]. ■

Lemma III.5. Suppose that the sequence $\{x_k\}_{k\geq 0}$ is generated by Algorithm 1 and Assumption 1 and 2 hold, then the Steps 3, 4, and 5 of Algorithm 1 are well-defined, which means that each step can be completed within the finite time in each iteration, regardless of the size of input data.

Proof: Set $\hat{d}_k^j$ be the approximate solution of the trust region subproblem II.9 at x_k for $j \in \mathbb{N}$, it is evident that x_k is not the optimal solution, thus we have $\|\hat{g}_k\| \geq \delta$. Then, according to $-\frac{g_x\left(\hat{g}_k, q_k\right)}{\left\|\hat{g}_k\right\| \cdot \left\|q_k\right\|} \geq \tau$, we obtain

$$-\frac{g_x\left(\hat{g}_{\dot{k}}, q_{\dot{k}}\right)}{\left\|q_{\dot{k}}\right\|} \geq \tau\delta. \tag{III.19}$$

Based on the equation III.19, Lemma III.2 and Lemma III.4, the equation satisfies

$$\begin{aligned} &\left|\frac{\hat{f}(x_k) - \hat{f}\left(R_{x_k}\left(\hat{d}_k^j\right)\right)}{\hat{m}\left(0\right) - \hat{m}\left(\hat{d}_k^j\right)} - 1\right| \\ &= \left|\frac{\hat{f}\left(x_{\dot{k}}\right) - \hat{f}\left(R_{x_{\dot{k}}}\left(\hat{d}_{\dot{k}}^j\right)\right) - \hat{m}\left(0\right) + \hat{m}\left(\hat{d}_{\dot{k}}^j\right)}{\hat{m}\left(0\right) - \hat{m}\left(\hat{d}_{\dot{k}}^j\right)}\right| \\ &\leq \frac{o\left(\left\|\hat{d}_k^j\right\|^2\right)}{\frac{1}{2}\rho_k^{\alpha j}\min\left\{\frac{1}{Q_1}\left(\frac{-g_x(\hat{g}_k, q_k)}{\|q_k\|}\right)^2, \kappa\left(\frac{-g_x(\hat{g}_k, q_k)}{\|q_k\|}\right)\right\}} \\ &\leq \frac{o\left(\rho_{\dot{\kappa}}^{\alpha j}\min\left\{s_{\dot{\kappa}}, \kappa\right\}^2\right)}{\frac{1}{2}\rho_{\dot{\kappa}}^{\alpha j}\min\left\{\frac{\left(\tau\delta\right)^2}{Q_1}, \kappa(\tau\delta)\right\}}. \end{aligned} \tag{III.20}$$

Since $\lim\limits_{j\to\infty}\left(\overline{\Delta}_k^j = \rho_k^{\alpha j}\min\left\{s_{\dot{k}}, \kappa\right\}\right) \to 0$, combing with the inequation III.20, we have

$$\lim_{j\to\infty}\frac{\hat{f}\left(x_k\right) - \hat{f}\left(R_{x_k}\left(\hat{d}_k^j\right)\right)}{\hat{m}\left(0\right) - \hat{m}\left(\hat{d}_k^j\right)} = 1. \tag{III.21}$$

Additionally, according to Lemma III.3, we can obtain

$$\hat{\rho}_k^j = \frac{D_k - \hat{f}\left(R_{x_k}\left(\hat{d}_k^j\right)\right)}{\hat{m}(0) - \hat{m}\left(\hat{d}_k^j\right)} \geq \frac{\hat{f}(x_k) - \hat{f}\left(R_{x_k}\left(\hat{d}_k^j\right)\right)}{\hat{m}(0) - \hat{m}\left(\hat{d}_k^j\right)}, \tag{III.22}$$

which implies that $\lim_{j\to\infty} \hat{\rho}_k^j \geq 1 > \mu \in (0,1)$, hence the step 5 are well-defined. ■

Theorem III.1. Suppose Assumption 1 and 2 hold true, then Algorithm 1 either halts at a fixed point or generates an infinite sequence $\{x_k\}_{k\geq 0}$ such that

$$\lim_{k\to\infty} -\frac{g_x(\hat{g}_k, q_k)}{\|q_k\|} = 0. \tag{III.23}$$

Proof: When Algorithm 1 does not halt at a fixed point, the equation III.23 should be true. Suppose that there exist constants $\varepsilon_0 > 0$ and an infinite subset $\theta \subseteq \mathbb{N} \cup \{0\}$ such that

$$-\frac{g_x(\hat{g}_k, q_k)}{\|q_k\|} \geq \varepsilon_0, k \in \theta, \tag{III.24}$$

by employing Lemma III.2 and equation III.24, we obtain

$$\begin{aligned}
&\hat{f}(x_k) - \hat{f}\left(R_{x_k}\left(\hat{d}_k\right)\right) \geq \mu\left(\hat{m}(0) - \hat{m}\left(\hat{d}_k\right)\right) \\
&\geq \frac{1}{2}\mu\rho^{\alpha}\min\left\{\frac{1}{Q_1}\left(\frac{-g_x(\hat{g}_k, q_k)}{\|q_k\|}\right)^2, \kappa\left(\frac{-g_x(\hat{g}_k, q_k)}{\|q_k\|}\right)\right\} \\
&\geq \frac{1}{2}\mu\rho^{\alpha}\min\left\{\frac{\varepsilon_0^2}{Q_1}, \kappa\varepsilon_0\right\}.
\end{aligned} \tag{III.25}$$

Suppose the sequence $\{x_k\}_{k\geq 0}$ enerated by Algorithm 1, then $\{x_k\}_{k\geq 0} \subseteq C(x_0)$. By utilizing $\hat{\rho}_k \geq \mu$, we have

$$\begin{aligned}
&\hat{f}(x_k) \geq \mu\left(\hat{m}(0) - \hat{m}\left(\hat{d}_k\right)\right) + \hat{f}\left(R_{x_k}\left(\hat{d}_k\right)\right) \\
&\geq \hat{f}\left(R_{x_k}\left(\hat{d}_k\right)\right),
\end{aligned} \tag{III.26}$$

which indicates that the sequence $\left\{\hat{f}(x_k)\right\}$ is monotonically decreasing. Moreover, according to Assumption 1, the sequence $\left\{\hat{f}(x_k)\right\}$ converges.

Thus, $\frac{1}{2}\mu\rho^{\alpha}\min\left\{\frac{\varepsilon_0^2}{Q_1}, \kappa\varepsilon_0\right\} \leq 0$, which contradicts the conclusion of Lemma III.5, that us, there does not exist an infinite subset $\theta \subseteq \mathbb{N} \cup \{0\}$ such that III.24 holds true. Overall, the conclusion is established. ■

Theorem III.2 (Global Convergence). Suppose that all the conditions of Theorem III.1 hold true, then Algorithm 1 either terminates in a finite number of steps, or it generates an infinite sequence $\{x_k\}_{k\geq 0}$ such that $\lim_{k\to\infty} \|\hat{g}_k\| = 0$.

Proof: Case 1: If Algorithm 1 terminates under finite iterations, then the conclusion obviously holds true.

Case 2: Assuming Algorithm 1 generates an infinite sequence $\{x_k\}_{k\geq 0}$, and q_k satisfies $\frac{-g_x(\hat{g}_k, q_k)}{\|\hat{g}_k\|\cdot\|q_k\|} \geq \tau$, sequence $\{x_k\}_{k\geq 0}$, and q_k satisfies the infinite sequence $\{x_k\}_{k\geq 0}$ satisfies III.23. When $k \to \infty$, we obtain

$$0 \leq \tau\|\hat{g}_k\| \leq -\frac{g_x(\hat{g}_k, q_k)}{\|\hat{g}_k\|\cdot\|q_k\|}\|\hat{g}_k\| = -\frac{g_x(\hat{g}_k, q_k)}{\|q_k\|} \to 0. \tag{III.27}$$

Therefore, the conclusion $\lim_{k\to\infty} \|\hat{g}_k\| = 0$ holds. ■

B. Local Convergence

In this section, we analyze the local convergence rate of Algorithm 1, which includes proving R-linear convergence and super-linear convergence. The proof methodology is based on references [38], [39].

Lemma III.6. Reference [24] suppose that f is a twice continuously differentiable function on M and is uniformly contractively convex, meaning that f satisfies convexity and contractivity properties on M. Suppose that

$$\frac{D}{dt}\left(\frac{d}{dt}R_x(t\xi)\right)\Big|_{t=0} = 0, \quad \forall x \in M, \xi \in T_xM, \tag{III.28}$$

where $\frac{D}{dt}$ represents the covariant derivative along the curve $t \mapsto R_x(t\xi)$.

Consider now the existence of $0 < p \leq q$ such that

$$p\|\xi\|^2 \leq g_x(Hessf(x)[\xi], \xi) \leq q\|\xi\|^2. \tag{III.29}$$

By Lemma III.1, we have

$$\begin{aligned}
&p\Big(d_R(y,x)\Big)^2 \leq g_x\Big(P_{\gamma}^{0\leftarrow 1} gradf(y) - gradf(x), d_R(y,x)\Big) \\
&\leq q\left(d_R(y,x)\right)^2, \quad \forall x, y \in M
\end{aligned} \tag{III.30}$$

and

$$p\Big(d_R\Big(x, x^+\Big)\Big) \leq \Big\|gradf(x)\Big\|, \quad \forall x \in M \tag{III.31}$$

where x^+ is a stationary point of f.

Theorem III.3 (R-linear Convergence). Suppose that all the assumptions and conclusions of Lemma III.6 hold true, as well as Assumption 1 and Assumption 2. And $\frac{d}{dt}R_{x_k}$ is equally continuous on a neighborhood δ_n of the stationary point x^+. There exists $L > 0$ such that for sufficiently large k, it holds that $L\Big\|\hat{g}_k\Big\| \leq \sqrt{g_x\Big(\hat{d}_k, \hat{d}_k\Big)}$. Then, the sequence $\{x_k\}_{k\geq 0}$ generated by Algorithm 1 converges R-linearly to the stationary point x^+.

Proof: According to Theorem III.2 and equation III.31, we have

$$\lim_{k\to\infty} p\Big(d_R\Big(x_k, x^+\Big)\Big) \leq \lim_{k\to\infty}\Big\|\hat{g}_k\Big\| = 0, \tag{III.32}$$

thus, $\lim_{k\to\infty} x_k = x^+$. By utilizing II.10 and II.12, we obtain

$$\begin{aligned}
&D_k - \hat{f}\Big(R_{x_k}\Big(\hat{d}_k\Big)\Big) \geq \mu\Big(\hat{m}(0) - \hat{m}\Big(\hat{d}_k\Big)\Big) \\
&\geq \mu\beta\|\hat{g}_k\|\min\left\{\overline{\Delta}_k, \frac{\|\hat{g}_k\|}{\|H_{x_k}\|}\right\} \\
&= \mu\beta\|\hat{g}_k\|\min\left\{\rho^{\alpha}\min\{s_k, \kappa\}, \frac{\|\hat{g}_k\|}{\|H_{x_k}\|}\right\}.
\end{aligned} \tag{III.33}$$

By the conclusion of Assumption 2 and $g_x\left(\hat{d}_{\dot{k}},\hat{d}_{\dot{k}}\right)\leq\left(\rho^{\alpha}\min\left\{s_{\dot{k}},\kappa\right\}\right)^2$, we have

$$\begin{aligned}&\hat{f}\Big(R_{x_k}\left(\hat{d}_k\right)\Big)\\&\leq D_k-\mu\beta\left\|\hat{g}_k\right\|\min\left\{\rho^{\alpha}\min\left\{s_k,\kappa\right\},\frac{\|\hat{g}_k\|}{\|H_{x_k}\|}\right\}\\&\leq D_{\dot{k}}-\mu\beta\left\|\hat{g}_{\dot{k}}\right\|^2\min\left\{\frac{\sqrt{g_x\left(\hat{d}_{\dot{k}},\hat{d}_{\dot{k}}\right)}}{\left\|\hat{g}_{\dot{k}}\right\|},\frac{1}{Q_1}\right\}\\&\leq D_{\dot{\kappa}}-\mu\beta\left\|\hat{g}_{\dot{\kappa}}\right\|^2\min\left\{L,\frac{1}{Q_1}\right\},\end{aligned}\tag{III.34}$$

set $\partial=\mu\beta\min\left\{L,\frac{1}{Q_1}\right\}$, then III.34 is equal to

$$\hat{f}\Big(R_{x_k}\Big(\hat{d}_k\Big)\Big)\leq D_k-\partial\Big\|\hat{g}_k\Big\|^2.\tag{III.35}$$

For $k\in\mathbb{N}$, there exists $\Re>0$ such that $\sqrt{g_x\left(\hat{d}_k,\hat{d}_k\right)}\leq\Re\|\hat{g}_k\|$, then

$$d\left(R_{x_k}\left(\hat{d}_k\right),x_k\right)=\sqrt{g_x\left(\hat{d}_k,\hat{d}_k\right)}\leq\Re\|\hat{g}_k\|,\tag{III.36}$$

according to Assumption 1 and equation III.36, there exists $\Re_1>0$ such that

$$\begin{aligned}&\|\hat{g}_{k+1}\|\leq\left\|\hat{g}_{k+1}-P_{\gamma}^{x_{k+1}\leftarrow x_k}\hat{g}_k\right\|+\|\hat{g}_k\|\\&\leq\Re_1 d\Big(x_k,x_kR_{x_k}\Big(\hat{d}_k\Big)\Big)+\|\hat{g}_k\|\\&=\Re_1\sqrt{g_x\left(\hat{d}_k,\hat{d}_k\right)}+\|\hat{g}_k\|\leq\left(L\Re_1+1\right)\|\hat{g}_k\|.\end{aligned}\tag{III.37}$$

We consider the validity of III.38, i.e.,

$$D_{k+1}-f\left(x^+\right)\leq T_P\left(D_k-f\left(x^+\right)\right),\tag{III.38}$$

where $T_P=\left(1-\partial\right)\in\left(0,1\right)$.

Suppose that $\left\|\hat{g}_k\right\|^2\leq D_k-f\left(x^+\right)$, by using II.11, we have

$$\begin{aligned}&D_{k+1}-f\Big(x^+\Big)\\&=\eta_{k+1}D_k+\left(1-\eta_{k+1}\right)\hat{f}\left(R_{x_k}\left(\hat{d}_k\right)\right)-f\left(x^+\right),\end{aligned}\tag{III.39}$$

according to the conclusion of Lemma III.3, we have

$$\begin{aligned}&D_{k+1}-f\left(x^+\right)\leq\eta_{k+1}\hat{f}\left(R_{x_k}\left(\hat{d}_k\right)\right)+\\&\left(1-\eta_{k+1}\right)\hat{f}\left(R_{x_k}\left(\hat{d}_k\right)\right)-f\left(x^+\right)\\&=\hat{f}\left(R_{x_k}\left(\hat{d}_k\right)\right)-f\left(x^+\right),\end{aligned}\tag{III.40}$$

by using III.35, we obtain

$$\begin{aligned}&D_{k+1}-f\left(\ x^+\ \right)\leq D_k-\partial\left\|\hat{g}_k\right\|^2-f\left(\ x^+\ \right)\\&\leq\left(1-\partial\right)\left(D_k-f\left(x^+\right)\right)\end{aligned}\tag{III.41}$$

and

$$\begin{aligned}&D_{k+1}-f\left(x^+\right)\leq T_P\left(D_k-f\left(x^+\right)\right)\\&\leq\cdots\leq T_P^{k-1}\left(D_0-f\left(x^+\right)\right).\end{aligned}\tag{III.42}$$

Furthermore, assume that condition III.43 holds, i.e.,

$$\begin{aligned}&\frac{p}{2}\Big(d_{_R}\Big(x,x^+\Big)\Big)^2\leq f(x)-f(x^+)\\&\leq\frac{q}{2}\Big(d_{_R}\Big(x,x^+\Big)\Big)^2,\quad\forall x\in M.\end{aligned}\tag{III.43}$$

Combining III.43 with Lemma III.3, we can deduce

$$\begin{aligned}&\left(d_R\left(x_{k+1},x^+\right)\right)^2\leq\frac{2}{p}\left(\hat{f}\left(R_{x_k}\left(\hat{d}_k\right)\right)-f\left(x^+\right)\right)\\&\leq\frac{2}{p}\Big(D_{k+1}-f\Big(x^+\Big)\Big)\leq\cdots\leq\frac{2}{p}T_P^{k+1}\Big(D_0-f\Big(x^+\Big)\Big)\\&=\frac{2}{p}T_P^{k+1}\left(\hat{f}\left(x_0\right)-f\left(x^+\right)\right)\end{aligned}\tag{III.44}$$

then,

$$\begin{aligned}&\lim_{k\to\infty}\Big(d\Big(x_{k+1},x^+\Big)\Big)^{\frac{1}{k+1}}\leq\lim_{k\to\infty}\Big(d_R\Big(x_{k+1},x^+\Big)\Big)^{\frac{1}{k+1}}\\&\leq\lim_{k\to\infty}\left(\sqrt{\frac{2}{p}T_P^{k+1}\left(\hat{f}\left(x_0\right)-f\left(x^+\right)\right)}\right)^{\frac{1}{k+1}}<1.\end{aligned}\tag{III.45}$$

This implies that the sequence $\{x_k\}_{k\geq0}$ generated by Algorithm 1 converges R-linearly to the stationary point x^+. ■

Theorem III.4 (Super-Linear Convergence). Suppose that all the assumptions and conclusions of Lemma III.1, Lemma III.6 hold true, as well as Assumption 1 and Assumption 2, H_{x_k} is symmetric positive definite. The sequence generated $\{x_k\}_{k\geq0}$ by Algorithm 1 converges to x^+, and $\frac{d}{dt}R_{x_k}$ is equally continuous on a neighborhood δ_n of the stationary point x^+, possessing

$$\begin{aligned}&\lim_{k\to\infty}\frac{\left\|H_{x_k}q_{\dot{k}}-P_{\gamma}^{x_k\leftarrow x^+}\left(Hessf\left(x^+\right)\left(P_{\gamma}^{x^+\leftarrow x_k}q_{\dot{k}}\right)\right)\right\|}{\left\|q_{\dot{k}}\right\|}\\&=0\end{aligned}\tag{III.46}$$

where $q_k=-\left(H_{x_k}\right)^{-1}\hat{g}_k$. Then, the sequence $\{x_k\}_{k\geq0}$ converges super-linearly to x^+, such that

$$\left\|R_{x_k}\left(\hat{d}_k\right)-x^+\right\|=o\left(x_k-x^+\right).\tag{III.47}$$

Proof: Due to $\overline{\Delta}_{\dot{\kappa}}=\rho^{\alpha}\min\left\{s_{\dot{\kappa}},\kappa\right\}$, Case 1: If $s_{\dot{\kappa}}\leq\kappa$, thus $\overline{\Delta}_{\dot{\kappa}}=\rho^{\alpha}s_{\dot{\kappa}}$, according to the definition of $s_k\big(k\geq1\big)$ in equation II.7:

$$\begin{aligned}&\overline{\Delta}_k=\rho^{\alpha}\max\left\{-\frac{g_x(\hat{g}_k,q_k)}{q_k^TH_{x_k}q_k}\|q_k\|,\lambda\hat{\Delta}_{k-1}\right\}\\&\geq-\rho^{\alpha}\frac{g_x(\hat{g}_{\dot{k}},q_{\dot{k}})}{q_{\dot{k}}^TH_{x_k}q_{\dot{k}}}\|q_{\dot{k}}\|.\end{aligned}\tag{III.48}$$

For $\alpha=0$, we have $\overline{\Delta}_{\dot{k}}\geq-\rho^{\alpha}\frac{g_x(\hat{g}_{\dot{k}},q_{\dot{k}})}{q_{\dot{k}}^TH_{x_k}q_{\dot{k}}}\|q_{\dot{k}}\|=\|q_{\dot{k}}\|$, which implies that the trust region subproblem II.9 has an approximate solution $\hat{d}_k^*=q_k$.In the following section, we will analyze the validity of III.49:

$$\frac{D_{\dot{k}}-\hat{f}\left(R_{x_k}\left(\hat{d}_k^*\right)\right)}{\hat{m}\left(0\right)-\hat{m}\left(\hat{d}_k^*\right)}\geq\mu,\quad\exists k_0\in\mathbb{N},\forall k\geq k_0,\tag{III.49}$$

according to $q_k = -\left(H_{x_k}\right)^{-1}\hat{g}_k$ and the conclusion $\|H_{x_k}\| \leq Q_1$ of Assumption 2, we have

$$\left\|\hat{d}_{\dot{k}}^*\right\| = \left\|q_{\dot{k}}\right\| = \left\|-\left(H_{x_k}\right)^{-1}\hat{g}_{\dot{k}}\right\| \leq \frac{1}{Q_1}\left\|\hat{g}_{\dot{k}}\right\|. \quad \text{(III.50)}$$

Equation III.32 indicates that $\lim\limits_{\dot{k}\to\infty}\|\hat{g}_{\dot{k}}\| = 0$, which implies $\lim_{\dot{\kappa}\to\infty}\left\|\hat{d}_{\dot{\kappa}}^*\right\| = 0$. Since $q_k = -\left(H_{x_k}\right)^{-1}\hat{g}_k$, it is evident that $-g_x\left(\hat{g}_k, q_k\right) = q_k^T H_{x_k} q_k$, according to the conclusion of III.5 and $\alpha = 0$, we obtain

$$\hat{m}(0) - \hat{m}\left(\hat{d}_k^*\right) \geq \frac{1}{2}\frac{\left(g_x\left(\hat{g}_k, q_k\right)\right)^2}{q_k^T H_{x_k} q_k} = \frac{1}{2}q_k^T H_{x_k} q_k, \quad \text{(III.51)}$$

simultaneously, Lemma III.4 implies

$$\left|\hat{f}\left(R_{x_k}\left(\hat{d}_k^*\right)\right) - \hat{f}\left(x_k\right) - \hat{m}\left(0\right) + \hat{m}\left(\hat{d}_k^*\right)\right| \leq o\left(\left\|\hat{d}_k^*\right\|^2\right). \quad \text{(III.52)}$$

Therefore, based on the above discussion, we have

$$\left|\frac{\hat{f}\left(x_k\right) - \hat{f}\left(R_{x_k}\left(\hat{d}_k^*\right)\right)}{\hat{m}\left(0\right) - \hat{m}\left(\hat{d}_k^*\right)} - 1\right| \leq \frac{2o\left(\left\|\hat{d}_k^*\right\|^2\right)}{q_k^T H_{x_k} q_k} \leq \frac{o\left(\left\|\hat{d}_k^*\right\|^2\right)}{Q_1\left\|\hat{d}_k^*\right\|^2}, \quad \text{(III.53)}$$

i.e., $\lim_{k\to\infty}\frac{\hat{f}(x_k) - \hat{f}\left(R_{x_k}\left(\hat{d}_k^*\right)\right)}{\hat{m}(0) - \hat{m}\left(\hat{d}_k^*\right)} = 1$. Then, considering $\exists k_0 \in \mathbb{N}, \forall k \geq k_0$, we obtain

$$\begin{aligned}&\frac{D_{\dot{\kappa}} - \hat{f}\left(R_{x_k}\left(\hat{d}_{\dot{\kappa}}^*\right)\right)}{\hat{m}\left(0\right) - \hat{m}\left(\hat{d}_{\dot{\kappa}}^*\right)} \geq \\ &\frac{\hat{f}\left(x_k\right) - \hat{f}\left(R_{x_k}\left(\hat{d}_{\dot{k}}^*\right)\right)}{\hat{m}\left(0\right) - \hat{m}\left(\hat{d}_{\dot{k}}^*\right)} \geq \mu \in (0,1).\end{aligned} \quad \text{(III.54)}$$

Case 2: If $s_{\dot{\kappa}} > \kappa$, then $\overline{\Delta}_{\dot{\kappa}} = \rho^{\alpha}\kappa$, set $\alpha = 0$, considering $-\frac{g_x(\hat{g}_k, q_k)}{q_k^T H_{x_k} q_k}\|q_k\| \leq \kappa$. Using the same proof technique as in Case 1, we can derive equation III.54, which implies that the sequence $\{x_k\}_{k\geq 0}$ converges super-linearly to the stationary point $x^{\mp}$, such that $\left\|R_{x_k}\left(\hat{d}_k\right) - x^+\right\| = o\left(x_k - x^+\right)$. ■

IV. Numerical Experiments

In this section, Independent Component Analysis (ICA) problem [40], [41] is applied to verify the effectiveness of the proposed Algorithm 1. To further illustrate its performance, we compare the numerical performance of Algorithm 1 with RTR [24] and NMRTR algorithm [38] in solving ICA problem. ICA is a mathematical technique which is mainly used to decompose a multivariate mixed signal into different source signals, which fundamental assumption is that the multivariate mixed signal is combined by the linear independent components. The goal of ICA is to decompose the multivariate mixed signal into its original source signals. In practical, the covariance matrix of the signals may not be completely diagonalized. To address this issue, a "soft-whitening" step is introduced to simultaneously diagonalize the source condition and the mixed covariance matrix, which can help enhance the stability and accuracy of ICA. The ICA problem is addressed by minimizing the objective function on the Stiefel manifold, shown as

$$\begin{aligned}&f_d(X){:}St(n,r) \to \mathbb{R} : X \mapsto f_d(X) \\ &= -\sum_{i=1}^{N}\left\|diag(X^T C_i X)\right\|_F^2,\end{aligned} \quad \text{(IV.1)}$$

where $\|diag(W)\|_F^2$ represents the sum of squares of the diagonal elements of W, and C_i is a symmetric cumulant or time-lagged covariance matrix. For $\varpi_1, \varpi_2 \in T_X St(n,r)$ the canonical inner product $g\left(\varpi_1, \varpi_2\right) = trace\left(\varpi_1^T \varpi_2\right)$ is used to define the Riemannian metric g on the manifold. The metric g is commonly referred to as the Frobenius inner product. The retraction is defined as

$$R_X : T_X St(n,r) \to St(n,r) : \varpi \mapsto R_X\varpi := hk\left(X + \varpi\right) \quad \text{(IV.2)}$$

where $hk(X)$ represents the matrix Q obtained by the QR decomposition of matrix X, and the tangent space of $St(n,r)$ at point X is given by the following formula,

$$T_X St(n,r) = \left\{Y \in \mathbb{R}^{n\times r} : X^T Y + Y^T X = 0\right\} \quad \text{(IV.3)}$$

In addition, the gradient of the objective function $f_d(X)$ is defined as [40], [41],

$$gradf_d(X) = P_X(-\sum_{i}^{N} 4C_i X diag(X^T C_i X)), \quad \text{(IV.4)}$$

where P_X denotes the orthogonal projection of X onto $T_X St(n,r)$, which is defined as,

$$P_X\left(\ \varpi\ \right) = \varpi - X sym\left(X^T\varpi\right), \quad \text{(IV.5)}$$

where $sym(K)$ represents a diagonal matrix, which diagonal elements is the diagonal elements of matrix K. And the initial Hessian matrix of $f_d(X)$ is given by the following

$$Hessf_d(X)[\varpi] = -4\sum_{i}^{N}\left(P_X\left(\varphi_1 + 2\varphi_2\right) - \varphi_3\right), \quad \text{(IV.6)}$$

where

$$\begin{aligned}\varphi_1 &= C_i\varpi diag\left(X^T C_i X\right), \\ \varphi_2 &= C_i X diag\left(X^T C_i \varpi\right), \\ \varphi_3 &= \varpi sym\left(X^T C_i X diag\left(X^T C_i X\right)\right),\end{aligned}$$

and $\varpi \in T_X St(n,r)$.

In Algorithm 1, the termination criterion is set as $\|\hat{g}_k\| \leq 10^{-6}\|\hat{g}_0\|$ or $k \geq \max k$, where the number of maximum iteration is no more than 30000. The settled parameters of Algorithm 1 are as follows $\hat{\Delta}_k = 0.8$, $\mu = 0.5$, $\tau = 0.7$, $\lambda = 1.5$, $\kappa = 1.0$. Furthermore, we choose a small-scale problem with dimension $(n, r, N) = (12, 4, 256)$. In the experiment, we mainly study the numerical performance of Algorithm 1 on a full Hessian matrix. Simultaneously, we set up three different LBFGS memory buffers and adjust the size of the memory buffer by setting the corresponding parameter m, which effectively governs the storage of historical information about the approximate Hessian matrix. Moreover, let n_i denote the number of iterations in gradient updating,

TABLE I: Numerical Comparison with Full Hessian in Algorithm

Algorithm	RTR	NMRTR	Our Algorithm
n_i	44	36	30
n_h	379	341	321
$n_f\ (s)$	22.68	19.53	17.82
iteration	44	36	30

n_h represent the number of iterations in Hessian matrix updating, and n_f indicate the average loading time of CPU.

The numerical experiments reveal that Algorithm 1 outperforms RTR and NMRTR algorithms in terms of the number of iterations in gradient updating, iterations in Hessian matrix updating, and average loading time of CPU. The results demonstrate our algorithm takes the better performance in solving small-scale ICA problem.

V. Conclusion

This paper introduces a novel approach, the Nonmonotone Adaptive Trust Region Algorithm on Manifolds(NATRAM), which demonstrates significant advantages over the RTR and NMRTR algorithms in addressing optimization problems with orthogonal constraints. We transform the optimization problem into an unconstrained optimization problem on the Stiefel manifold, enabling unconstrained optimization algorithms on manifolds can be applied to solve the problem. According to the proof of convergence, we have demonstrated that NATRAM exhibits both global and local convergence properties. And the numerical experiments indicate that NATRAM takes advantages in solving small-scale ICA problems. In the future, we will further validate the performance of the proposed method in a broader range of optimization problems with orthogonal constraints.

Acknowledgment

The authors wish to thank Xianfeng Ding, Xiaolin Yi, Hanbing Mei, Yiyu Qin, and Qianmei Luo for useful discussions.

References

[1] P.-A. Absil, R. Mahony, and R. Sepulchre, Optimization algorithms on matrix manifolds. Princeton University Press, 2008.

[2] A. Edelman, T. A. Arias, and S. T. Smith, "The geometry of algorithms with orthogonality constraints," SIAM journal on Matrix Analysis and Applications, vol. 20, no. 2, pp. 303–353, 1998.

[3] J. Li, L. Fuxin, and S. Todorovic, "Efficient riemannian optimization on the stiefel manifold via the cayley transform," arXiv preprint arXiv:2002.01113, 2020.

[4] Q. Wang and W. H. Yang, "Proximal quasi-newton method for composite optimization over the stiefel manifold," Journal of Scientific Computing, vol. 95, no. 2, p. 39, 2023.

[5] X. Liu, Z. Wen, X. Wang, M. Ulbrich, and Y. Yuan, "On the analysis of the discretized kohn–sham density functional theory," SIAM Journal on Numerical Analysis, vol. 53, no. 4, pp. 1758–1785, 2015.

[6] M. Mrovec and J. Berger, "A diagonalization-free optimization algorithm for solving kohn–sham equations of closed-shell molecules," Journal of Computational Chemistry, vol. 42, no. 7, pp. 492–504, 2021.

[7] C. Yang, J. C. Meza, and L.-W. Wang, "A trust region direct constrained minimization algorithm for the kohn–sham equation," SIAM Journal on Scientific Computing, vol. 29, no. 5, pp. 1854–1875, 2007.

[8] L. Eldén and H. Park, "A procrustes problem on the stiefel manifold," Numerische Mathematik, vol. 82, no. 4, pp. 599–619, 1999.

[9] I. Söderkvist, "Perturbation analysis of the orthogonal procrustes problem," BIT Numerical Mathematics, vol. 33, pp. 687–694, 1993.

[10] A. d'Aspremont, L. Ghaoui, M. Jordan, and G. Lanckriet, "A direct formulation for sparse pca using semidefinite programming," Advances in neural information processing systems, vol. 17, 2004.

[11] W. W. Hager, D. T. Phan, and J. Zhu, "Projection algorithms for nonconvex minimization with application to sparse principal component analysis," Journal of Global Optimization, vol. 65, pp. 657–676, 2016.

[12] Z. Wen, C. Yang, X. Liu, and Y. Zhang, "Trace-penalty minimization for large-scale eigenspace computation," Journal of Scientific Computing, vol. 66, pp. 1175–1203, 2016.

[13] H. Sato and T. Iwai, "A riemannian optimization approach to the matrix singular value decomposition," SIAM Journal on Optimization, vol. 23, no. 1, pp. 188–212, 2013.

[14] B. Gao, X. Liu, and Y.-x. Yuan, "Parallelizable algorithms for optimization problems with orthogonality constraints," SIAM Journal on Scientific Computing, vol. 41, no. 3, pp. A1949–A1983, 2019.

[15] Z. Wen and W. Yin, "A feasible method for optimization with orthogonality constraints," Mathematical Programming, vol. 142, no. 1, pp. 397–434, 2013.

[16] W. Huang, P.-A. Absil, and K. A. Gallivan, "A riemannian symmetric rank-one trust-region method," Mathematical Programming, vol. 150, no. 2, pp. 179–216, 2015.

[17] P.-A. Absil, C. G. Baker, and K. A. Gallivan, "Trust-region methods on riemannian manifolds with applications in numerical linear algebra," in Proceedings of the 16th International Symposium on Mathematical Theory of Networks and Systems (MTNS2004), Leuven, Belgium. Citeseer, 2004, pp. 5–9.

[18] S. J. Wright, "Numerical optimization," 2006.

[19] M. Berger, "A panoramic view of riemannian geometry," 2003.

[20] N. Boumal, An introduction to optimization on smooth manifolds. Cambridge University Press, 2023.

[21] W. M. Boothby, An introduction to differentiable manifolds and Riemannian geometry. Academic press, 1986.

[22] H. Sato, Riemannian optimization and its applications. Springer, 2021, vol. 670.

[23] M. P. Do Carmo and F. Francis, "J. riemannian geometry. vol. 6," 1992.

[24] P.-A. Absil, C. G. Baker, and K. A. Gallivan, "Trust-region methods on riemannian manifolds," Foundations of Computational Mathematics, vol. 7, pp. 303–330, 2007.

[25] J. H. Manton, "Optimization algorithms exploiting unitary constraints," IEEE transactions on signal processing, vol. 50, no. 3, pp. 635–650, 2002.

[26] W. Huang, "Optimization algorithms on riemannian manifolds with applications," Ph.D. dissertation, The Florida State University, 2013.

[27] W. Ring and B. Wirth, "Optimization methods on riemannian manifolds and their application to shape space," SIAM Journal on Optimization, vol. 22, no. 2, pp. 596–627, 2012.

[28] W. Huang, K. A. Gallivan, and P.-A. Absil, "A broyden class of quasi-newton methods for riemannian optimization," SIAM Journal on Optimization, vol. 25, no. 3, pp. 1660–1685, 2015.

[29] 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 Journal of Numerical Analysis, vol. 22, no. 3, pp. 359–390, 2002.

[30] H. Kasai and B. Mishra, "Inexact trust-region algorithms on riemannian manifolds," Advances in neural information processing systems, vol. 31, 2018.

[31] N. Eslami, B. Najafi, and S. M. Vaezpour, "A trust region method for solving multicriteria optimization problems on riemannian manifolds," Journal of Optimization Theory and Applications, vol. 196, no. 1, pp. 212–239, 2023.

[32] N.-z. Gu and J.-t. Mo, "Incorporating nonmonotone strategies into the trust region method for unconstrained optimization," Computers & Mathematics with Applications, vol. 55, no. 9, pp. 2158–2172, 2008.

[33] Y. Xue, H. Liu, and Z. Liu, "An improved nonmonotone adaptive trust region method," Applications of Mathematics, vol. 64, no. 3, pp. 335–350, 2019.

TABLE II: Numerical Comparison of Algorithm with Different LBFGS Memory Buffers

Algorithm	RTR			NMRTR			Our Algorithm		
	m=2	m=5	m=10	m=2	m=5	m=10	m=2	m=5	m=10
n_i	413	504	672	389	466	601	365	437	552
n_h	0	0	0	0	0	0	0	0	0
$n_f\,(s)$	3.23	4.95	6.54	3.03	4.58	5.62	2.86	4.26	5.49
iteration	413	504	672	389	466	601	365	437	552

[34] A. Kamandi and K. Amini, "A new nonmonotone adaptive trust region algorithm," Applications of Mathematics, vol. 67, no. 2, pp. 233–250, 2022.

[35] K. Amini and M. Ahookhosh, "A hybrid of adjustable trust-region and nonmonotone algorithms for unconstrained optimization," Applied Mathematical Modelling, vol. 38, no. 9-10, pp. 2601–2612, 2014.

[36] N. Boumal, P.-A. Absil, and C. Cartis, "Global rates of convergence for nonconvex optimization on manifolds," IMA Journal of Numerical Analysis, vol. 39, no. 1, pp. 1–33, 2019.

[37] A. Kamandi, K. Amini, and M. Ahookhosh, "An improved adaptive trust-region algorithm," Optimization Letters, vol. 11, pp. 555–569, 2017.

[38] X. Li, X. Wang, and M. Krishan Lal, "A nonmonotone trust region method for unconstrained optimization problems on riemannian manifolds," Journal of Optimization Theory and Applications, vol. 188, pp. 547–570, 2021.

[39] X.-b. Li, N.-j. Huang, Q. H. Ansari, and J.-C. Yao, "Convergence rate of descent method with new inexact line-search on riemannian manifolds," Journal of Optimization Theory and Applications, vol. 180, pp. 830–854, 2019.

[40] F. J. Theis, T. P. Cason, and P. A. Absil, "Soft dimension reduction for ica by joint diagonalization on the stiefel manifold," in Independent Component Analysis and Signal Separation: 8th International Conference, ICA 2009, Paraty, Brazil, March 15-18, 2009. Proceedings 8. Springer, 2009, pp. 354–361.

[41] P.-A. Absil and K. A. Gallivan, "Joint diagonalization on the oblique manifold for independent component analysis," in 2006 IEEE international conference on acoustics speech and signal processing proceedings, vol. 5. IEEE, 2006, pp. V–V.