

Riemannian Stochastic Variance-Reduced Cubic Regularized Newton Method for Submanifold Optimization

Dewei Zhang and Sam Davanloo Tajbakhsh*

{zhang.8705,davanloo.1}@osu.edu

The Ohio State University

July 25, 2022

Abstract

We propose a stochastic variance-reduced cubic regularized Newton algorithm to optimize the finite-sum problem over a Riemannian submanifold of the Euclidean space. The proposed algorithm requires a full gradient and Hessian update at the beginning of each epoch while it performs stochastic variance-reduced updates in the iterations within each epoch. The iteration complexity of $O(\epsilon^{-3/2})$ to obtain an $(\epsilon, \sqrt{\epsilon})$ -second-order stationary point, i.e., a point with the Riemannian gradient norm upper bounded by ϵ and minimum eigenvalue of Riemannian Hessian lower bounded by $-\sqrt{\epsilon}$, is established when the manifold is embedded in the Euclidean space. Furthermore, the paper proposes a computationally more appealing modification of the algorithm which only requires an *inexact* solution of the cubic regularized Newton subproblem with the same iteration complexity. The proposed algorithm is evaluated and compared with three other Riemannian second-order methods over two numerical studies on estimating the inverse scale matrix of the multivariate t-distribution on the manifold of symmetric positive definite matrices and estimating the parameter of a linear classifier on the Sphere manifold.

Keywords— Riemannian optimization, manifold optimization, stochastic optimization, cubic regularization, variance reduction.

1 Introduction

We study the optimization of the finite-sum problem over a Riemannian manifold \mathcal{M} embedded in a Euclidean space \mathcal{E} as

$$\min_{\mathbf{x} \in \mathcal{M} \subseteq \mathcal{E}} F(\mathbf{x}) = \frac{1}{N} \sum_{i=1}^N f_i(\mathbf{x}), \quad (1)$$

where N is a (possibly very large) positive integer. Manifold optimization has a range of applications in machine learning, statistics, control and robotics, e.g., in deep learning, low-rank matrix completion, sparse or nonnegative principal component analysis, or solving large-scale semidefinite programs – see Hu *et al.* (2020), Absil & Hosseini (2019) and the references therein. The finite-sum structure of the objective function in problem (1) specifically finds applications in machine

*Corresponding author

learning and statistics for parameter estimation, and addition of the manifold constraint could have problem-specific, computational, or other reasons. Below, we provide two motivational examples for problem (1).

Example 1 (Parameter estimation of the multivariate Student’s t-distribution)

As an important member of the family of elliptical distributions Domino (2018), the multivariate t-distribution has numerous applications in mathematical finance, survival analysis, biology, etc. Kotz & Nadarajah (2004). For instance in mathematical finance Szegö (2002), de Melo Mendes & de Souza (2004), Krzanowski & FHC (1994), the *Student’s t copula* $C_\nu : [0, 1]^p \rightarrow [0, 1]$ defined as $C_\nu(\mathbf{u}) = T_{\nu, \Sigma, \boldsymbol{\mu}}(T_\nu^{-1}(u_1), \dots, T_\nu^{-1}(u_n))$, where $T_{\nu, \Sigma, \boldsymbol{\mu}}(\mathbf{x})$ is the multivariate Student’s t cumulative density function (CDF) and $T_\nu^{-1}(\cdot)$ is the inverse of the marginal univariate Student’s t CDF is used to model or sample from multivariate Student’s t-distribution Krzanowski & FHC (1994). As one of the core tasks, the maximum likelihood parameter estimation of the multivariate Student’s t-distribution requires solving

$$\max_{\Sigma \in \mathcal{S}_{++}^p, \boldsymbol{\mu} \in \mathbb{R}^p} \frac{1}{N} \sum_{i=1}^N \log(t_\nu(\mathbf{x}_i; \boldsymbol{\mu}, \Sigma)), \quad (2)$$

where $t_\nu(\mathbf{x}; \boldsymbol{\mu}, \Sigma)$ denotes the probability density function of the multivariate t-distribution and $\nu \in \mathbb{N}_+$ is the degrees of freedom which is generally predetermined. Since the scale matrix Σ should belong to the manifold of positive-definite matrices, problem (2) is an instance of problem (1).

Example 2 (Efficient training of deep neural networks)

Training deep neural networks could be “notoriously difficult” when the singular values of the hidden-to-hidden weight matrices deviates from one Arjovsky *et al.* (2016). In such cases optimization becomes difficult due to the *vanishing* or *exploding* gradient Arjovsky *et al.* (2016), Wisdom *et al.* (2016). This challenge can be circumvented, if the weight matrices are unitary with singular values equal to one. This can be achieved by requiring hidden-to-hidden weight matrices to belong to Stiefel Manifold $St(p, n) \triangleq \{X \in \mathbb{R}^{n \times p} : X^\top X = I_p\}$ Absil *et al.* (2009), Boumal (2020) and training the model using a Riemannian optimization algorithm. The underlying optimization problem is an instance of (1) over the cartesian product of Stiefel manifolds. The orthonormality of the weight matrices improves the performance of deep neural networks Bansal *et al.* (2018), Li *et al.* (2020), reduces overfitting to improve generalization Cogswell *et al.* (2015), or stabilizes the distribution of activations over layers Huang *et al.* (2018). In Sun *et al.* (2017) and Xie *et al.* (2017) two convolutional neural networks are trained with orthonormal weight matrices for phase retrieval and image classification.

1.1 Related Work

Numerous algorithms for standard unconstrained optimization Ruszczynski (2011) have been generalized to Riemannian manifolds Absil *et al.* (2009), Udriste (2013), Boumal (2020). Some notable first-order algorithms include gradient descent method Zhang & Sra (2016), Boumal *et al.* (2019), conjugate gradient method Smith (1994), Sato & Iwai (2015), stochastic gradient method Bonnabel (2013), Zhang *et al.* (2016), Tripuraneni *et al.* (2018), accelerated methods Liu *et al.* (2017), Ahn & Sra (2020), Zhang & Sra (2018), Criscitiello & Boumal (2020), Alimisis *et al.* (2021, 2020), proximal gradient methods Ferreira & Oliveira (2002), Bento *et al.* (2015, 2017), de Carvalho Bento *et al.* (2016), Huang & Wei (2021). To guarantee convergence to a second-order stationary point, Jin *et al.* (2019) investigates the Riemannian perturbed gradient descent that guarantees second-order stationarity without using second-order information with $\tilde{O}(1/\epsilon^2)$ iteration complexity. Other

saddle-escape methods over manifolds were also studied in Sun *et al.* (2019) and Criscitiello & Boumal (2019).

In the context of second-order algorithms, the Newton method is extended to optimize over Riemannian manifolds in Luenberger (1972), Gabay (1982), Smith (1993, 1994). Specifically, Smith (1993) and Smith (1994) establish local quadratic convergence. Similar to the Newton method on Euclidean space, the Newton method for manifold optimization also suffers from two main drawbacks: first, it is possible that the Hessian matrix is degenerated at a point; second, it is possible that the iterates diverge, converge to a saddle point, or even a local maximum. In the Riemannian setting, the trust region method Absil *et al.* (2004, 2007), Baker *et al.* (2008), Boumal (2015), Boumal *et al.* (2019) and the cubic regularized Newton method Zhang & Zhang (2018) are extensions of their Euclidean counterparts to address these drawbacks. More specifically, Boumal *et al.* (2019) shows that the Riemannian trust region method obtains an (ϵ, ϵ) -second-order-stationary point (see Definition 4.2) in $O(1/\epsilon^3)$ which matches its Euclidean counterpart Cartis *et al.* (2012, 2014). Furthermore, the cubic regularized Newton method on manifolds Zhang & Zhang (2018) is shown to reach an $(\epsilon, \sqrt{\epsilon})$ -second-order stationary point in $O(1/\epsilon^{1.5})$. Finally, Agarwal *et al.* (2018) extends the daptive cubic regularization method Cartis *et al.* (2011a) to Riemannian manifolds and establishes $O(1/\epsilon^{1.5})$ rate to obtain an $(\epsilon, \sqrt{\epsilon})$ -second-order stationary point – see also Qi (2011), Hu *et al.* (2018).

A common issue among second-order algorithms is their high computational cost to calculate the inverse of the Hessian operator. A Riemannian counterpart of the famous BFGS algorithm Nocedal & Wright (2006) is proposed in Ring & Wirth (2012) which does not require calculating the inverse of the Hessian operator.

To optimize functions with the finite-sum structure or those that are known through their approximate gradient and Hessian, inexact methods including first- and second-order Riemannian stochastic algorithms and their variance-reduced extensions are proposed in the literature. As a generalization of Johnson & Zhang (2013), the Riemannian stochastic variance-reduced gradient descent (SVRG) method was developed in Zhang *et al.* (2016). Furthermore, an extension of Riemannian SVRG with computationally more efficient retraction and vector transport was developed in Sato *et al.* (2019). The paper also establishes global convergence properties of their method besides its local convergence rate. The Riemannian version of the stochastic recursive gradient method Nguyen *et al.* (2017) is proposed in Kasai *et al.* (2018). Kasai & Mishra (2018) proposes Riemannian trust region algorithms with inexact gradient and Hessian that allows inexact solution of the subproblem. Furthermore, a Riemannian stochastic variance-reduce quasi-Newton method is proposed in Kasai *et al.* (2017) - see also Roychowdhury (2017). For a recent review of first- and second-order Riemannian optimization algorithms, we refer the reader to Hosseini & Sra (2020) and Sato (2021) (specifically, Section 6.1 on stochastic methods).

1.2 Contributions

The major contributions of this paper are as follows: (i) Motivated by Zhou *et al.* (2018) and Kovalev *et al.* (2019) in the Euclidean setting, we propose a stochastic variance-reduced cubic regularized Newton method (R-SVRC algorithm) to optimize over Riemannian manifolds. (ii) We carefully analyze the worst-case complexity of the proposed algorithm to find a point that satisfy the first- and second-order necessary optimality conditions (i.e., a second-order stationarity point) when the cubic-regularized Newton subproblem is solved *exactly* - see Theorem 4.1 and Corollary 4.1. (iii) We performed the analysis of a computationally more appealing version of the algorithm, that

allows solving the cubic-regularized Newton subproblem *inexactly*, and established the same worst-case complexity bound - see Theorem 4.2 and Corollary 4.2. The assumptions for our analysis are explicitly discussed in Section 4. Finally, the performance of the proposed algorithm is evaluated and compared over two applications: 1. Estimating the scale matrix of Student's t-distribution over the symmetric positive definite manifold, 2. Learning the parameter of a linear classifier over a Sphere manifold. The implementation of the proposed algorithm in MATLAB with exact and inexact subproblem solvers is provided at <https://github.com/samdavanloo/R-SVRC>. To the best of our knowledge, this work is the first *stochastic* Newton method with cubic regularization on Riemannian manifold.

1.3 Preliminaries and Notation

A Riemannian manifold (\mathcal{M}, g) is a real smooth manifold \mathcal{M} equipped with a Riemannian metric g . The metric g induces an inner product structure in each tangent space $T_{\mathbf{x}}\mathcal{M}$ associated with point $\mathbf{x} \in \mathcal{M}$. We denote the inner product of $\mathbf{u}, \mathbf{v} \in T_{\mathbf{x}}\mathcal{M}$ as $\langle \mathbf{u}, \mathbf{v} \rangle_{\mathbf{x}} = g_{\mathbf{x}}(\mathbf{u}, \mathbf{v})$, and the norm of \mathbf{u} is defined as $\|\mathbf{u}\| = \sqrt{g_{\mathbf{x}}(\mathbf{u}, \mathbf{u})}$. Furthermore, the angle between \mathbf{u} and \mathbf{v} is $\arccos(\langle \mathbf{u}, \mathbf{v} \rangle_{\mathbf{x}} / (\|\mathbf{u}\| \|\mathbf{v}\|))$. Given a smooth real-valued function f on a Riemannian manifold \mathcal{M} , Riemannian gradient and Hessian of f at \mathbf{x} are denoted by $\text{grad}f(\mathbf{x})$ and $\text{Hess}f(\mathbf{x})$ (also for simplicity by $H_{\mathbf{x}}$). For a symmetric operator, e.g. the Riemannian Hessian H at $\mathbf{x} \in \mathcal{M}$, the operator norm of H is defined as $\|H\|_{op} = \sup\{\|H\eta\| : \eta \in T_{\mathbf{x}}\mathcal{M}, \|\eta\| = 1\}$. An operator on $T_{\mathbf{x}}\mathcal{M}$ is positive semidefinite $H \succeq 0$ if $\langle H[\eta], \eta \rangle \geq 0$, for any $\eta \in T_{\mathbf{x}}\mathcal{M}$. A geodesic is a constant speed curve $\gamma : [0, 1] \rightarrow \mathcal{M}$ that is locally distance minimizing. An exponential map $\text{Exp}_{\mathbf{x}} : T_{\mathbf{x}}\mathcal{M} \rightarrow \mathcal{M}$ maps $\mathbf{v} \in T_{\mathbf{x}}\mathcal{M}$ to $\mathbf{y} \in \mathcal{M}$, such that there is a geodesic γ with $\gamma(0) = \mathbf{x}$, $\gamma(1) = \mathbf{y}$, and $\dot{\gamma}(0) = \mathbf{v}$. For two points $\mathbf{x}, \mathbf{y} \in \mathcal{M}$, and $d(\mathbf{x}, \mathbf{y}) < \text{inj}(\mathcal{M})$, there is a unique geodesic. The exponential map has an inverse $\text{Exp}_{\mathbf{x}}^{-1} : \mathcal{M} \rightarrow T_{\mathbf{x}}\mathcal{M}$ and the geodesic is the unique shortest path with $\|\text{Exp}_{\mathbf{x}}^{-1}(\mathbf{y})\| = \|\text{Exp}_{\mathbf{y}}^{-1}(\mathbf{x})\|$ the geodesic distance between $\mathbf{x}, \mathbf{y} \in \mathcal{M}$. Parallel transport $\Gamma_{\mathbf{x}}^{\mathbf{y}} : T_{\mathbf{x}}\mathcal{M} \rightarrow T_{\mathbf{y}}\mathcal{M}$ maps a vector $\mathbf{v} \in T_{\mathbf{x}}\mathcal{M}$ to $\Gamma_{\mathbf{x}}^{\mathbf{y}}\mathbf{v} \in T_{\mathbf{y}}\mathcal{M}$, while preserving norm, and roughly speaking "direction". A tangent vector of a geodesic γ remains tangent if parallel transported along γ . Parallel transport also preserves inner products, i.e. $\langle \mathbf{u}, \mathbf{v} \rangle_{\mathbf{x}} = \langle \Gamma(\gamma)_{\mathbf{x}}^{\mathbf{y}}\mathbf{u}, \Gamma(\gamma)_{\mathbf{x}}^{\mathbf{y}}\mathbf{v} \rangle_{\mathbf{y}}$. We denote the orthogonal projection operator onto $T_{\mathbf{x}}\mathcal{M}$ by $P_{\mathbf{x}}$.

Let (\mathcal{M}, g) be a connected Riemannian manifold (see e.g. Absil *et al.* (2009)) which carries the structure of a metric space whose distance function is the arc length of a minimizing path between two points.

Definition 1.1 (Riemannian metric). *An inner product on $T_{\mathbf{x}}\mathcal{M}$ is a bilinear, symmetric, positive definite function $\langle \cdot, \cdot \rangle_{\mathbf{x}} : T_{\mathbf{x}}\mathcal{M} \times T_{\mathbf{x}}\mathcal{M} \rightarrow \mathbb{R}$. It induces a norm for tangent vectors as $\|u\|_{\mathbf{x}} = \sqrt{\langle u, u \rangle_{\mathbf{x}}}$. The smoothly varying inner product is called the Riemannian metric, i.e., if \mathbf{v}, \mathbf{w} are two smooth vector fields on \mathcal{M} then the function $\mathbf{x} \mapsto \langle \mathbf{v}(\mathbf{x}), \mathbf{w}(\mathbf{x}) \rangle_{\mathbf{x}}$ is smooth from \mathcal{M} to \mathbb{R} .*

Remark 1. *The inner product of two elements $\xi_{\mathbf{x}}$ and $\zeta_{\mathbf{x}}$ of $T_{\mathbf{x}}\mathcal{M}$ are interchangeably denoted by $g(\xi_{\mathbf{x}}, \zeta_{\mathbf{x}}) = g_{\mathbf{x}}(\xi_{\mathbf{x}}, \zeta_{\mathbf{x}}) = \langle \xi_{\mathbf{x}}, \zeta_{\mathbf{x}} \rangle = \langle \xi_{\mathbf{x}}, \zeta_{\mathbf{x}} \rangle_{\mathbf{x}}$.*

Definition 1.2 (Injectivity radius Boumal (2020)). *The injectivity radius of a Riemannian manifold \mathcal{M} at a point \mathbf{x} , denoted by $\text{inj}(\mathbf{x})$, is the supremum over radius $r > 0$ such that $\text{Exp}_{\mathbf{x}}$ is defined and is a diffeomorphism on the open ball $B(\mathbf{x}, r) = \{v \in T_{\mathbf{x}}\mathcal{M} : \|v\|_{\mathbf{x}} < r\}$. By the inverse function theorem, $\text{inj}(\mathbf{x}) > 0$. Furthermore, the injectivity radius of a Riemannian manifold \mathcal{M} , i.e., $\text{inj}(\mathcal{M})$, is the infimum of $\text{inj}(\mathbf{x})$ over $\mathbf{x} \in \mathcal{M}$ (Boumal (2020), Definition 10.14).*

Consider the ball $U \triangleq B(\mathbf{x}, \text{inj}(\mathbf{x})) \subseteq T_{\mathbf{x}}\mathcal{M}$ in the tangent space at \mathbf{x} . Its image $\mathcal{U} \triangleq \text{Exp}_{\mathbf{x}}(U)$ is a neighborhood of \mathbf{x} in \mathcal{M} . By definition, $\text{Exp}_{\mathbf{x}} : U \rightarrow \mathcal{U}$ is a diffeomorphism, with well-defined, smooth inverse $\text{Exp}_{\mathbf{x}}^{-1} : \mathcal{U} \rightarrow U$. With these choices of domains, $v = \text{Exp}_{\mathbf{x}}^{-1}(\mathbf{y})$ is the unique shortest tangent vector at \mathbf{x} such that $\text{Exp}_{\mathbf{x}}(v) = \mathbf{y}$.

Definition 1.3 (Riemannian distance). *The Riemannian distance on a connected Riemannian manifold (\mathcal{M}, g) is*

$$d : \mathcal{M} \times \mathcal{M} \rightarrow \mathbb{R} : d(\mathbf{x}, \mathbf{y}) \triangleq \inf_{\gamma \in \Gamma} L(\gamma), \quad (3)$$

where $L(\gamma) = \int_a^b \sqrt{g_{\gamma(t)}(\dot{\gamma}(t), \dot{\gamma}(t))} dt$ and Γ is the set of all curves in \mathcal{M} joining points \mathbf{x} and \mathbf{y} . Specifically, if $\|v\|_{\mathbf{x}} < \text{inj}(\mathbf{x})$ then $d(\mathbf{x}, \text{Exp}_{\mathbf{x}}(v)) = \|v\|_{\mathbf{x}}$.

Definition 1.4 (Riemannian gradient). *Given a smooth real-valued function f on a Riemannian manifold \mathcal{M} , the Riemannian gradient of f at a point $\mathbf{x} \in \mathcal{M}$, denoted by $\text{grad}f(\mathbf{x})$, is defined as the unique element of $T_{\mathbf{x}}\mathcal{M}$ that satisfies*

$$\langle \text{grad}f(\mathbf{x}), \xi \rangle_{\mathbf{x}} = Df(\mathbf{x})[\xi], \quad \forall \xi \in T_{\mathbf{x}}\mathcal{M}. \quad (4)$$

Specifically, when \mathcal{M} is a Riemannian submanifold of the Euclidean space $\mathbb{R}^{m \times n}$,

$$\text{grad}f(\mathbf{x}) = P_{\mathbf{x}} \nabla f(\mathbf{x}), \quad (5)$$

where $P_{\mathbf{x}}$ is the Euclidean projection onto $T_{\mathbf{x}}\mathcal{M}$ which is a nonexpansive linear transformation.

Definition 1.5 (Riemannian Hessian). *Given a real-valued function f on a Riemannian manifold \mathcal{M} , the Riemannian Hessian of f at a point $\mathbf{x} \in \mathcal{M}$ is the linear mapping $\text{Hess}f(\mathbf{x})$ from $T_{\mathbf{x}}\mathcal{M}$ onto itself defined as*

$$\text{Hess}f(\mathbf{x})[\xi] = \nabla_{\xi} \text{grad}f \quad (6)$$

for all $\xi \in \mathcal{M}$, where ∇ is the Riemannian connection on \mathcal{M} Absil et al. (2009).

When \mathcal{M} is a Riemannian submanifold of the Euclidean space $\mathbb{R}^{m \times n}$, the Riemannian Hessian of f is written as

$$\text{Hess}f(\mathbf{x})[\xi] = P_{\mathbf{x}}(D\text{grad}f(\mathbf{x})[\xi]), \quad (7)$$

i.e., the classical directional derivatives followed by an orthogonal projection. For more information, e.g., refer to Proposition 5.3.2 in Absil et al. (2009).

2 Proposed Algorithm

The proposed Riemannian Stochastic Variance-Reduced Cubic Regularization (R-SVRC) method is presented in Algorithm 1. This algorithm is indeed semi-stochastic, which requires calculation of full gradient and Hessian at the beginning of each epoch s , i.e., the outer loop in the algorithm. However, within each epoch, there are T iterations of the inner loop, which require calculation of stochastic variance-reduced gradient and Hessian by sampling $|I_g|$ and $|I_h|$ components, respectively.

From the computational perspective, the major step of the algorithm is to solve the cubic-regularized Newton subproblem. Under Assumption 3, the manifold is embedded in $\mathbb{R}^{m \times n}$; hence, the tangent vectors in $T_{\mathbf{x}}\mathcal{M}$ are naturally represented by $m \times n$ matrices (see Absil et al. (2009)). Therefore, current solvers for cubic regularized problem in Euclidean space can be adopted Boumal et al. (2014). Solving this generally nonconvex subproblem is discussed in more details below (27).

For computational gain, the paper also considers solving the subproblem inexactly. As long as the inexact solution satisfies the conditions in Definition 4.4, our analysis guarantees the results of the exact case.

Algorithm 1 Riemannian Stochastic Variance-Reduced Cubic Regularization (R-SVRC)

Require: batch size parameters b_g, b_h , cubic penalty parameters σ , number of epochs S , epoch length T , and a starting point \mathbf{x}_0 .

- 1: Set $\hat{\mathbf{x}}^1 = \mathbf{x}_0$
 - 2: **for** $s = 1, \dots, S$ **do**
 - 3: $\mathbf{x}_0^s = \hat{\mathbf{x}}^s$
 - 4: $\mathbf{g}^s = \text{grad}F(\hat{\mathbf{x}}^s) = \frac{1}{N} \sum_{i=1}^N \text{grad}f_i(\hat{\mathbf{x}}^s)$; $\mathbf{H}^s = \text{Hess}F(\hat{\mathbf{x}}^s) = \frac{1}{N} \sum_{i=1}^N \text{Hess}f_i(\hat{\mathbf{x}}^s)$
 - 5: **for** $t = 0, \dots, T-1$ **do**
 - 6: Sample index set I_g, I_h , s.t. $|I_g| = b_g, |I_h| = b_h$
 - 7: Compute $\hat{\eta}_t^s \in T_{\hat{\mathbf{x}}^s}$, s.t. $\text{Exp}_{\hat{\mathbf{x}}^s}(\hat{\eta}_t^s) = \mathbf{x}_t^s$
 - 8: $\mathbf{v}_t^s = \Gamma_{\hat{\mathbf{x}}^s}^{\mathbf{x}_t^s}(\mathbf{g}^s) + \frac{1}{b_g} (\sum_{i_t \in I_g} \text{grad}f_{i_t}(\mathbf{x}_t^s) - \Gamma_{\hat{\mathbf{x}}^s}^{\mathbf{x}_t^s}(\sum_{i_t \in I_g} \text{grad}f_{i_t}(\hat{\mathbf{x}}^s))) - \Gamma_{\hat{\mathbf{x}}^s}^{\mathbf{x}_t^s}(\frac{1}{b_g} \sum_{i_t \in I_g} \text{Hess}f_{i_t}(\hat{\mathbf{x}}^s) - \mathbf{H}^s) \hat{\eta}_t^s$
 - 9: $\mathbf{h}_t^s = \text{argmin}_{\mathbf{h} \in T_{\mathbf{x}} \mathcal{M}} \langle \mathbf{v}_t^s, \mathbf{h} \rangle + \frac{1}{2} \langle \mathbf{U}_t^s \mathbf{h}, \mathbf{h} \rangle + \frac{\sigma}{6} \|\mathbf{h}\|^3$, where
 $\mathbf{U}_t^s = \Gamma_{\hat{\mathbf{x}}^s}^{\mathbf{x}_t^s} \circ \mathbf{H}^s \circ \Gamma_{\mathbf{x}_t^s}^{\hat{\mathbf{x}}^s} + \frac{1}{b_h} \sum_{j_t \in I_h} \text{Hess}f_{j_t}(\mathbf{x}_t^s) - \frac{1}{b_h} \Gamma_{\hat{\mathbf{x}}^s}^{\mathbf{x}_t^s} \circ (\sum_{j_t \in I_h} \text{Hess}f_{j_t}(\hat{\mathbf{x}}^s)) \circ \Gamma_{\mathbf{x}_t^s}^{\hat{\mathbf{x}}^s}$
 - 10: $\mathbf{x}_{t+1}^s = \text{Exp}_{\mathbf{x}_t^s}(\mathbf{h}_t^s)$
 - 11: **end for**
 - 12: $\hat{\mathbf{x}}^{s+1} = \mathbf{x}_T^{s+1}$
 - 13: **end for**
 - 14: **return** $\mathbf{x}_{out} = \mathbf{x}_t^s$, where s, t are uniformly at random chosen from $s \in [S]$ and $t \in [T]$
-

3 Lipschitzian Smoothness on Riemannian Manifolds

Definition 3.1 (g-smoothness da Cruz Neto *et al.* (1998), Ferreira *et al.* (2019)). *A differentiable function $f : \mathcal{M} \rightarrow \mathbb{R}$ is said to be geodesically L_g -smooth if its gradient is L_g -Lipschitz, i.e., for any $\mathbf{x}, \mathbf{y} \in \mathcal{M}$ with $d(\mathbf{x}, \mathbf{y}) < \text{inj}(\mathcal{M})$,*

$$\|\text{grad}f(\mathbf{x}) - \Gamma_{\mathbf{y}}^{\mathbf{x}} \text{grad}f(\mathbf{y})\|_{\mathbf{x}} \leq L_g d(\mathbf{x}, \mathbf{y}), \quad (8)$$

where $\Gamma_{\mathbf{y}}^{\mathbf{x}}$ is the parallel transport from \mathbf{y} to \mathbf{x} following the unique minimizing geodesic connecting \mathbf{x} and \mathbf{y} .

It can be proven that if f is L_g -smooth, then for any $\mathbf{x}, \mathbf{y} \in \mathcal{M}$ with $d(\mathbf{x}, \mathbf{y}) < \text{inj}(\mathcal{M})$,

$$|f(\mathbf{y}) - (f(\mathbf{x}) + \langle \text{Exp}_{\mathbf{x}}^{-1}(\mathbf{y}), \text{grad}f(\mathbf{x}) \rangle_{\mathbf{x}})| \leq \frac{L_g}{2} d^2(\mathbf{x}, \mathbf{y}) \quad (9)$$

– see, e.g. Bento *et al.* (2017), Lemma 2.1.

Definition 3.2 (H-smoothness Agarwal *et al.* (2020)). *A twice differentiable function $f : \mathcal{M} \rightarrow \mathbb{R}$ is said to be geodesically L_H -smooth if its Hessian is L_H -Lipschitz, i.e., for any $\mathbf{x}, \mathbf{y} \in \mathcal{M}$ with $d(\mathbf{x}, \mathbf{y}) < \text{inj}(\mathcal{M})$,*

$$\|\text{Hess}f(\mathbf{y}) - \Gamma_{\mathbf{x}}^{\mathbf{y}} \text{Hess}f(\mathbf{x}) \Gamma_{\mathbf{y}}^{\mathbf{x}}\|_{op} \leq L_H d(\mathbf{x}, \mathbf{y}). \quad (10)$$

It is shown in the following lemma that if f is L_H -smooth, then for any $\mathbf{x}, \mathbf{y} \in \mathcal{M}$ with $d(\mathbf{x}, \mathbf{y}) < \text{inj}(\mathcal{M})$, we have

$$|f(\mathbf{y}) - (f(\mathbf{x}) + \langle \eta, \text{grad}f(\mathbf{x}) \rangle_{\mathbf{x}} + \frac{1}{2} \langle \eta, \text{Hess}f(\mathbf{x})[\eta] \rangle_{\mathbf{x}})| \leq \frac{L_H}{6} d^3(\mathbf{x}, \mathbf{y}) \quad (11)$$

and

$$\|\text{grad}f(\mathbf{y}) - \Gamma_{\mathbf{x}}^{\mathbf{y}} \text{grad}f(\mathbf{x}) - \Gamma_{\mathbf{x}}^{\mathbf{y}} \text{Hess}f(\mathbf{x})\eta\|_{\mathbf{y}} \leq \frac{L_H}{2} d^2(\mathbf{x}, \mathbf{y}), \quad (12)$$

where $\eta = \text{Exp}_{\mathbf{x}}^{-1}(\mathbf{y})$.

Lemma 3.1 (Agarwal *et al.* (2020), Proposition 3.2). *If f is H -smooth with constant L_H , then (11) and (12) hold.*

The Lipschitz-type conditions above are parallel to the conditions in the Euclidean setting Nesterov & Polyak (2006). In general, it is not trivial to verify these conditions, or even determine their parameters. However, we know there is a broad class of functions on Euclidean space, which satisfy the Lipschitz continuity-related conditions. We conjectured similar properties as the Euclidean setting would imply (10), if \mathcal{M} is embedded in the Euclidean space. In Absil *et al.* (2009), it was proven that if the manifold is compact and the function has Lipschitz continuous gradient, then (8) holds. Boumal *et al.* (2019) proved that if the manifold is compact and the function has Lipschitz continuous gradient and Hessian, then (9) and (11) hold. In the following lemma, it is shown that (10) holds under the same conditions.

Lemma 3.2. *If \mathcal{M} is a compact submanifold of the Euclidean space \mathcal{E} and $f(\mathbf{x})$ has Lipschitz continuous Hessian in \mathcal{E} in the Euclidean sense, then (10) is satisfied.*

Proof. Denote the orthogonal projection operator onto $T_{\mathbf{x}}\mathcal{M}$, i.e. the tangent space of \mathcal{M} at \mathbf{x} , by $P_{\mathbf{x}}$. Denote the Euclidean gradient and Hessian by $\nabla f(\mathbf{x})$ and $\nabla^2 f(\mathbf{x})$ correspondingly. For any \mathbf{y} , such that $d(\mathbf{x}, \mathbf{y}) < \text{inj}(\mathcal{M})$ and any $\xi \in T_{\mathbf{y}}\mathcal{M}$, s.t. $\|\xi\| = 1$, we have

$$\begin{aligned} \text{Hess}f(\mathbf{y})[\xi] &= P_{\mathbf{y}}(D(\mathbf{y} \rightarrow P_{\mathbf{y}}\nabla f(\mathbf{y}))(\mathbf{y})[\xi]) \\ &= P_{\mathbf{y}}(D(\mathbf{y} \rightarrow P_{\mathbf{y}})(\mathbf{y})[\xi][\nabla f(\mathbf{y})]) + P_{\mathbf{y}}(\nabla^2 f(\mathbf{y})[\xi]) \\ &\equiv A_1 + B_1, \end{aligned}$$

The first equality follows from (7) and the second equality comes from the chain rule and the fact that the projection operator is linear. Similarly, we have

$$\begin{aligned} \Gamma_{\mathbf{x}}^{\mathbf{y}} \text{Hess}f(\mathbf{x})[\Gamma_{\mathbf{y}}^{\mathbf{x}}\xi] &= \Gamma_{\mathbf{x}}^{\mathbf{y}} P_{\mathbf{x}}(D(\mathbf{x} \rightarrow P_{\mathbf{x}}\nabla f(\mathbf{x}))(\mathbf{x})[\Gamma_{\mathbf{y}}^{\mathbf{x}}\xi]) \\ &= \Gamma_{\mathbf{x}}^{\mathbf{y}} P_{\mathbf{x}}(D(\mathbf{x} \rightarrow P_{\mathbf{x}})(\mathbf{x})[\Gamma_{\mathbf{y}}^{\mathbf{x}}\xi][\nabla f(\mathbf{x})]) + \Gamma_{\mathbf{x}}^{\mathbf{y}} P_{\mathbf{x}}(\nabla^2 f(\mathbf{x})[\Gamma_{\mathbf{y}}^{\mathbf{x}}\xi]) \\ &\equiv A_2 + B_2. \end{aligned}$$

First, to quantify $\|A_1 - A_2\|$, we have,

$$\|A_1 - A_2\| = \|O_{A_1}[\nabla f(\mathbf{y}) + \nabla f(\mathbf{x}) - \nabla f(\mathbf{x})] - O_{A_2}[\nabla f(\mathbf{x})]\| \quad (13)$$

$$= \|O_{A_1}[\nabla f(\mathbf{y}) - \nabla f(\mathbf{x})] + (O_{A_1} - O_{A_2})[\nabla f(\mathbf{x})]\| \quad (14)$$

$$\leq \|O_{A_1}[\nabla f(\mathbf{y}) - \nabla f(\mathbf{x})]\| + \|(O_{A_1} - O_{A_2})[\nabla f(\mathbf{x})]\| \quad (15)$$

$$\leq \|O_{A_1}\|_{op} \cdot \|\nabla f(\mathbf{y}) - \nabla f(\mathbf{x})\| + \|(O_{A_1} - O_{A_2})[\nabla f(\mathbf{x})]\| \quad (16)$$

where $O_{A_1} \triangleq P_{\mathbf{y}}(D(\mathbf{y} \rightarrow P_{\mathbf{y}})(\mathbf{y})[\xi][\cdot])$ and $O_{A_2} \triangleq \Gamma_{\tilde{\mathbf{x}}}^{\mathbf{y}} P_{\mathbf{x}}(D(\mathbf{x} \rightarrow P_{\mathbf{x}})(\mathbf{x})[\Gamma_{\tilde{\mathbf{y}}}^{\mathbf{x}} \xi][\cdot])$.

Due to the smoothness and compactness of \mathcal{M} and $\|\xi\| = 1$, $\|P_{\mathbf{y}}(D(\mathbf{y} \rightarrow P_{\mathbf{y}})(\mathbf{y})[\xi][\cdot])\|_{op}$ exists and is uniformly upper bounded, i.e. there exists a finite M_1 independent of \mathbf{x} , \mathbf{y} and ξ , s.t. $\|P_{\mathbf{y}}(D(\mathbf{y} \rightarrow P_{\mathbf{y}})(\mathbf{y})[\xi][\cdot])\|_{op} \leq M_1$ for any \mathbf{x} , $\mathbf{y} \in \mathcal{M}$ and ξ , s.t. $\|\xi\| = 1$.

For any \mathbf{z} , such that $d(\mathbf{z}, \mathbf{y}) < \text{inj}(\mathcal{M})$, define $Q_{\mathbf{z}, \mathbf{y}, \xi} \triangleq \Gamma_{\mathbf{z}}^{\mathbf{y}} P_{\mathbf{z}}(D(\mathbf{z} \rightarrow P_{\mathbf{z}})(\mathbf{z})[\Gamma_{\tilde{\mathbf{y}}}^{\mathbf{z}} \xi][\cdot])$. Note that $O_{A_1} = Q_{\mathbf{y}, \mathbf{y}, \xi}$ and $O_{A_2} = Q_{\mathbf{x}, \mathbf{y}, \xi}$. For fixed $\tilde{\mathbf{x}}$, $\tilde{\mathbf{y}}$ and $\tilde{\xi}$, $Q_{\mathbf{z}, \tilde{\mathbf{y}}, \tilde{\xi}}[\nabla f(\tilde{\mathbf{x}})]$ is a continuously differentiable function of \mathbf{z} based on the conditions that the manifold is smooth and $f(\mathbf{x})$ has Lipschitz continuous Hessian. Since \mathbf{z} belongs to a compact set, $Q_{\mathbf{z}, \tilde{\mathbf{y}}, \tilde{\xi}}[\nabla f(\tilde{\mathbf{x}})]$ is Lipschitz continuous on \mathbf{z} , i.e.

$$\|Q_{\mathbf{x}, \tilde{\mathbf{y}}, \tilde{\xi}}[\nabla f(\tilde{\mathbf{x}})] - Q_{\mathbf{y}, \tilde{\mathbf{y}}, \tilde{\xi}}[\nabla f(\tilde{\mathbf{x}})]\| \leq M_{\tilde{\mathbf{x}}, \tilde{\mathbf{y}}, \tilde{\xi}} \|\mathbf{x} - \mathbf{y}\|, \forall \mathbf{x}, \mathbf{y} \in \mathcal{M} \quad (17)$$

where $M_{\tilde{\mathbf{x}}, \tilde{\mathbf{y}}, \tilde{\xi}}$ is a finite constant depending on $\tilde{\mathbf{x}}, \tilde{\mathbf{y}}, \tilde{\xi}$. Especially, due to the smoothness of manifold and the function $f(\mathbf{x})$ has Lipschitz continuous Hessian, we have a continuous mapping from $\tilde{\mathbf{x}}, \tilde{\mathbf{y}}, \tilde{\xi}$ to $M_{\tilde{\mathbf{x}}, \tilde{\mathbf{y}}, \tilde{\xi}}$. Since $\tilde{\mathbf{x}}, \tilde{\mathbf{y}} \in \mathcal{M}$, which is a compact set and $\|\tilde{\xi}\| = 1$, we have a finite constant M_2 , s.t. $M_{\tilde{\mathbf{x}}, \tilde{\mathbf{y}}, \tilde{\xi}} \leq M_2$ for all $\tilde{\mathbf{x}}, \tilde{\mathbf{y}}, \tilde{\xi}$. In (17), letting $\mathbf{x} = \tilde{\mathbf{x}}$, $\mathbf{y} = \tilde{\mathbf{y}}$, we have,

$$\|Q_{\tilde{\mathbf{x}}, \tilde{\mathbf{y}}, \tilde{\xi}}[\nabla f(\tilde{\mathbf{x}})] - Q_{\tilde{\mathbf{y}}, \tilde{\mathbf{y}}, \tilde{\xi}}[\nabla f(\tilde{\mathbf{x}})]\| \leq M_{\tilde{\mathbf{x}}, \tilde{\mathbf{y}}, \tilde{\xi}} \|\tilde{\mathbf{x}} - \tilde{\mathbf{y}}\| \leq M_2 \|\tilde{\mathbf{x}} - \tilde{\mathbf{y}}\|. \quad (18)$$

Due to the arbitrariness of $\tilde{\mathbf{x}}$, $\tilde{\mathbf{y}}$ and $\tilde{\xi}$, we conclude the second term in (16), $\|(O_{A_1} - O_{A_2})[\nabla f(\mathbf{x})]\| \leq M_2 \|\mathbf{x} - \mathbf{y}\|$.

On the other hand, the gradient of a twice continuously differentiable function on a compact manifold is Lipschitz continuous. Therefore, there exists a finite L_1 , s.t.

$$\|A_1 - A_2\| \leq M_1 \|\nabla f(\mathbf{y}) - \nabla f(\mathbf{x})\| + M_2 \|\mathbf{x} - \mathbf{y}\| \leq (M_1 \cdot L_1 + M_2) \|\mathbf{y} - \mathbf{x}\| \leq (M_1 \cdot L_1 + M_2) d(\mathbf{x}, \mathbf{y}), \quad (19)$$

where $d(\mathbf{x}, \mathbf{y})$ is the Riemannian distance between \mathbf{x} and \mathbf{y} . The third inequality holds since the manifold is embedded in the Euclidean space.

Second, to quantify $\|B_1 - B_2\|$, we define

$$R_{\mathbf{y}, \xi}(\mathbf{z}) \triangleq \Gamma_{\mathbf{z}}^{\mathbf{y}} P_{\mathbf{z}}(\nabla^2 f(\mathbf{z})[\Gamma_{\tilde{\mathbf{y}}}^{\mathbf{z}} \xi]). \quad (20)$$

Fixing \mathbf{y}, ξ to be $\tilde{\mathbf{y}}$ and $\tilde{\xi}$, $R_{\tilde{\mathbf{y}}, \tilde{\xi}}(\mathbf{z})$ is Lipschitz continuous on \mathbf{z} due to the smoothness of the manifold and $\nabla^2 f(\mathbf{z})$ is Lipschitz continuous. Therefore, there exists a constant $N_{\tilde{\mathbf{y}}, \tilde{\xi}}$ depending on $\tilde{\mathbf{y}}$ and $\tilde{\xi}$, s.t. $\|R_{\tilde{\mathbf{y}}, \tilde{\xi}}(\mathbf{x}) - R_{\tilde{\mathbf{y}}, \tilde{\xi}}(\mathbf{y})\| \leq N_{\tilde{\mathbf{y}}, \tilde{\xi}} \|\mathbf{x} - \mathbf{y}\|$ for all $\mathbf{x}, \mathbf{y} \in \mathcal{M}$. Especially, there is a continuous mapping from $\tilde{\mathbf{y}}, \tilde{\xi}$ to $N_{\tilde{\mathbf{y}}, \tilde{\xi}}$. Since $\tilde{\mathbf{y}}, \tilde{\xi}$ are from compact sets, there exists a finite constant M_3 , s.t. $\|R_{\tilde{\mathbf{y}}, \tilde{\xi}}(\mathbf{x}) - R_{\tilde{\mathbf{y}}, \tilde{\xi}}(\mathbf{y})\| \leq N_{\tilde{\mathbf{y}}, \tilde{\xi}} \|\mathbf{x} - \mathbf{y}\| \leq M_3 \|\mathbf{x} - \mathbf{y}\|$ for all $\mathbf{x}, \mathbf{y} \in \mathcal{M}$.

Letting $\mathbf{x} = \tilde{\mathbf{x}}$, $\mathbf{y} = \tilde{\mathbf{y}}$, and due to the arbitrariness of $\tilde{\mathbf{x}}, \tilde{\mathbf{y}}$ and $\tilde{\xi}$, we have

$$\|B_1 - B_2\| = \|R_{\mathbf{y}, \xi}(\mathbf{y}) - R_{\mathbf{y}, \xi}(\mathbf{x})\| \leq M_3 \|\mathbf{x} - \mathbf{y}\| \leq M_3 \cdot d(\mathbf{x}, \mathbf{y}). \quad (21)$$

Combining (19), (21), there exists a finite $L \triangleq M_1 \cdot L_1 + M_2 + M_3$, s.t.

$$\|\text{Hess}f(\mathbf{y})[\xi] - \Gamma_{\mathbf{x}}^{\mathbf{y}}\text{Hess}f(\mathbf{x})[\Gamma_{\mathbf{y}}^{\mathbf{x}}\xi]\| \leq \|A_1 - A_2\| + \|B_1 - B_2\| \leq L \cdot d(\mathbf{x}, \mathbf{y})$$

Since ξ is an arbitrary tangent vector, we have,

$$\|\text{Hess}f(\mathbf{y}) - \Gamma_{\mathbf{x}}^{\mathbf{y}}\text{Hess}f(\mathbf{x})\Gamma_{\mathbf{y}}^{\mathbf{x}}\|_{op} \leq L \cdot d(\mathbf{x}, \mathbf{y}).$$

□

4 Complexity Analysis of the Proposed Algorithm

Definition 4.1 (Optimal gap). *For function $F(\cdot)$ and the initial point $\mathbf{x}_0 \in \mathcal{M}$, define*

$$\Delta_F \triangleq F(\mathbf{x}_0) - F^*, \quad (22)$$

where $F^* = \inf_{\mathbf{x} \in \mathcal{M}} F(\mathbf{x})$.

Without loss of generality, we assume $\Delta_F < +\infty$ throughout this paper.

Definition 4.2 ((ϵ, δ) -second-order stationary point). *\mathbf{x} is a second-order stationary point of the function $F : \mathcal{M} \rightarrow \mathbb{R}$ if $\|\text{grad}F(\mathbf{x})\| \leq \epsilon$ and $\lambda_{\min}(\text{Hess}F(\mathbf{x})) \geq -\delta$ where $\text{grad}F(\mathbf{x})$ and $\text{Hess}F(\mathbf{x})$ are the Riemannian gradient and Hessian of F at \mathbf{x} and $\lambda_{\min}(\text{Hess}F(\mathbf{x})) \triangleq \inf_{\eta \in T_{\mathbf{x}}\mathcal{M}} \{\langle \text{Hess}F(\mathbf{x})\eta, \eta \rangle_{\mathbf{x}} / \|\eta\|^2\}$.*

As in Nesterov & Polyak (2006), we define

$$\mu(\mathbf{x}) \triangleq \max\{\|\text{grad}F(\mathbf{x})\|^{3/2}, -\frac{\lambda_{\min}^3(\text{Hess}F(\mathbf{x}))}{L_H^{3/2}}\}. \quad (23)$$

In particular, according to the definition (23), $\mu(\mathbf{x}) \leq \epsilon^{3/2}$ holds if and only if

$$\|\text{grad}F(\mathbf{x})\| \leq \epsilon, \quad \lambda_{\min}(\text{Hess}F(\mathbf{x})) \geq -\sqrt{L_H\epsilon}. \quad (24)$$

Therefore, in order to find an $(\epsilon, \sqrt{L_H\epsilon})$ - approximate local minimum of the function defined over \mathcal{M} , it suffices to find $\mathbf{x} \in \mathcal{M}$ such that $\mu(\mathbf{x}) \leq \epsilon^{3/2}$.

Assumption 1. *We assume that the objective function F is bounded below, and its components f_i , $i = 1, \dots, N$ are twice continuously differentiable and they are g - and H -smooth.*

Assumption 2. *We assume that either i) functions f_i , $i = 1, \dots, N$ are Lipschitz continuous, or ii) functions f_i , $i = 1, \dots, N$ are continuously differentiable and the manifold \mathcal{M} is compact.*

Remark 2. *The g -smoothness of f_i , $i = 1, \dots, N$, in Assumption 1 implies that $\|\text{Hess}F(\mathbf{x})\|_{op}$ is bounded. Furthermore, Assumption 2 implies that $\|\text{grad}F(\mathbf{x})\|$ is bounded either by Lipschitz continuity of f_i , $i = 1, \dots, N$, or by the Weierstrass theorem Rudin et al. (1964). Hence, under Assumptions 1 and 2, and based on the fact that the parallel transport is isometric, there exist two positive constants c_g and c_H , such that*

$$\|\mathbf{v}_t^s\| \leq c_g \text{ and } \|U_t^s\|_{op} \leq c_H. \quad (25)$$

While the above two assumptions are mainly related to the objective function, the following three assumptions are related to the manifold.

Assumption 3. We assume that \mathcal{M} is embedded in a vector space, e.g., Euclidean space. For the ease of presentation, we assume $\mathcal{M} \subseteq \mathbb{R}^{m \times n}$.

Remark 3. Under Assumption 3, the Riemannian metric $g_{\mathbf{x}}(\cdot, \cdot)$ on the tangent space $T_{\mathbf{x}}\mathcal{M}$ is the restriction of the Euclidean metric. The norm induced by the Riemannian metric $\|\cdot\|_{\mathbf{x}}$ is the Euclidean norm.

Assumption 4. We assume that the manifold has positive injectivity radius, i.e. $\text{inj}(\mathcal{M}) \in (0, \infty]$ -see Definition 1.2.

Remark 4. To provide few examples, the unit sphere has injectivity radius equal to π , Hadamard manifolds and Euclidean spaces have infinite injectivity radius, and compact Riemannian manifolds have positive injectivity radius Chavel (2006). Note that the Assumption 4 implies that the manifold is complete.

Assumption 5. We assume the sectional curvature of the Riemannian manifold \mathcal{M} is lower-bounded by κ - see Lee (2018) for the definition of the sectional curvature.

Remark 5. Some manifolds that satisfy Assumption 5 include rotation group, hyperbolic manifold, the sphere, orthogonal groups, real projective space, Grassmann manifold, Stiefel manifold and compact subsets of the cone of positive definite matrices (see Bonnabel (2013), Sra & Hosseini (2015), Boumal (2020)).

Following the literature of the Newton method with cubic regularization Nesterov & Polyak (2006), Cartis *et al.* (2011a), we define

$$\tilde{m}(\mathbf{h}) = \langle \text{grad} f, \mathbf{h} \rangle + \frac{1}{2} \langle \text{Hess} f[\mathbf{h}], \mathbf{h} \rangle + \frac{\sigma}{6} \|\mathbf{h}\|^3, \mathbf{h} \in T_{\mathbf{x}}\mathcal{M}, \quad (26)$$

which can be regarded as a cubic regularization of locally quadratic approximation of function f - see Agarwal *et al.* (2018). From (11), we have $f(\text{Exp}_{\mathbf{x}}(\eta)) \leq \tilde{m}(\eta)$, for $\forall \eta \in T_{\mathbf{x}}\mathcal{M}$, if $\sigma \geq L_H$. From (26), we define

$$\mathbf{h}_t^s = \underset{\mathbf{h} \in T_{\mathbf{x}}\mathcal{M}}{\text{argmin}} m_t^s(\mathbf{h}), \quad (27)$$

where

$$m_t^s(\mathbf{h}) \triangleq \langle \mathbf{v}_t^s, \mathbf{h} \rangle + \frac{1}{2} \langle \mathbf{U}_t^s[\mathbf{h}], \mathbf{h} \rangle + \frac{\sigma}{6} \|\mathbf{h}\|^3 \quad (28)$$

and \mathbf{v}_t^s and \mathbf{U}_t^s are the approximated Riemannian gradient and Hessian operator of the objective function. Generally, (27) and (26) are not convex problems. Nesterov & Polyak (2006) proposed a way to transform these subproblems into convex programs in one variable. Recently, results in Carmon & Duchi (2019) show that under mild conditions gradient descent approximately finds the *global minimum* with the rate of $O(\epsilon^{-1} \log(1/\epsilon))$. Cartis *et al.* (2011b) propose a Lanczos-based method to minimize (26) exactly. The gradient, conjugate gradient and Newton methods to minimize (26) are available in the software package provided in Boumal *et al.* (2014).

We first provide some preliminary lemmas. Lemma 4.1 provides three identities that are used in the proofs of following lemmas. These identities are typical in the cubic regularization literature Nesterov & Polyak (2006), Cartis *et al.* (2011a,b). Lemma 4.2 provides an upper bound on

$\|\mathbf{h}_t^s\|$ which then provides a (lower) bound on the cubic regularization parameter σ to have the iterates close enough to the epoch points. Finally, Lemmas 4.3 and 4.4 provide upper bounds on the norm difference of $\text{grad}F$ and $\text{Hess}F$ with their variance-reduced estimators, and on the inner products $\langle \text{grad}F(\mathbf{x}_t^s) - \mathbf{v}_t^s, \eta \rangle$ and $\langle (\text{Hess}F(\mathbf{x}_t^s) - U_t^s)[\eta], \eta \rangle$ for $\eta \in T_{\mathbf{x}_t^s}\mathcal{M}$ which are used in the proofs of the main theorems.

Lemma 4.1. *Under Assumptions 3, for the semi-stochastic gradient and Hessian, we have*

$$\mathbf{v}_t^s + \mathbf{U}_t^s \mathbf{h}_t^s + \frac{\sigma}{2} \|\mathbf{h}_t^s\| \mathbf{h}_t^s = 0, \quad (29)$$

$$\mathbf{U}_t^s + \frac{\sigma}{2} \|\mathbf{h}_t^s\| \mathbf{I} \succeq 0, \quad (30)$$

$$\langle \mathbf{v}_t^s, \mathbf{h}_t^s \rangle + \frac{1}{2} \langle \mathbf{U}_t^s \mathbf{h}_t^s, \mathbf{h}_t^s \rangle + \frac{\sigma}{6} \|\mathbf{h}_t^s\|^3 \leq -\frac{\sigma}{12} \|\mathbf{h}_t^s\|^3. \quad (31)$$

Proof. (sketch) Under Assumption 3, i.e. the manifold is embedded in the Euclidean space, then the tangent space $T_{\mathbf{x}}\mathcal{M}$ in (27) is isomorphic to subspace of the Euclidean space. Hence, the proof follows, e.g., from that of Lemma 24 in Zhou *et al.* (2018). Indeed, the proof of (29) directly follows from the first-order optimality condition for a stationary point of (27). The inequality (30) relies on the fact that \mathbf{h}_t^s is a global minimizer which will not hold when solving (27) inexactly. The proof of (31) is based on (30) and (29). \square

Lemma 4.2. *Under Assumptions 1-4, given a constant $C > 0$ and $\sigma > \frac{2(c_g + C \cdot c_H)}{C^2}$, we have $\|\mathbf{h}_t^s\| < C$.*

Proof. Multiplying both sides of (29) by \mathbf{h}_t^s , we obtain $\langle \mathbf{v}_t^s, \mathbf{h}_t^s \rangle + \langle \mathbf{U}_t^s \mathbf{h}_t^s, \mathbf{h}_t^s \rangle + \frac{\sigma}{2} \|\mathbf{h}_t^s\|^3 = 0$. By Cauchy-Schwarz inequality, we have $\frac{\sigma}{2} \|\mathbf{h}_t^s\|^3 \leq \|\mathbf{v}_t^s\| \cdot \|\mathbf{h}_t^s\| + \|\mathbf{U}_t^s\|_{op} \cdot \|\mathbf{h}_t^s\|^2$. Dividing both sides by $\|\mathbf{h}_t^s\|$ and based on (25), we have $\frac{\sigma}{2} \|\mathbf{h}_t^s\|^2 - c_H \cdot \|\mathbf{h}_t^s\| - c_g \leq 0$, which implies

$$\|\mathbf{h}_t^s\| \leq \frac{c_H + \sqrt{c_H^2 + 2\sigma c_g}}{\sigma}. \quad (32)$$

Note that the right hand side of (32) is a monotonic decreasing function on σ . Hence, if $\sigma > \frac{2(c_g + C \cdot c_H)}{C^2}$, the right hand side of (32) is upper bounded by C , which implies $\|\mathbf{h}_t^s\| < C$. \square

Remark 6. *In Lemma 4.2 as well as Lemma 4.12, we set $C = \frac{\text{inj}(\mathcal{M})}{T}$, where T is the epoch length in Algorithm 1. Then, for any epoch s and iteration $t \in \{0, \dots, T-1\}$, we have*

$$d(\hat{\mathbf{x}}^s, \mathbf{x}_t^s) \leq \sum_{i=1}^t d(\mathbf{x}_{i-1}^s, \mathbf{x}_i^s) \leq \sum_{i=1}^t \|\mathbf{h}_i^s\| < \text{inj}(\mathcal{M}). \quad (33)$$

This inequality guarantees line 7 in Algorithm 1 is attained. In the following, we assume σ is large enough such that

$$\sigma > \frac{2(c_g T^2 + \text{inj}(\mathcal{M}) c_H T)}{(\text{inj}(\mathcal{M}))^2}, \quad (34)$$

hence, the distance between the iterate \mathbf{x}_t^s and $\hat{\mathbf{x}}^s$ is smaller than $\text{inj}(\mathcal{M})$.

In the proof of Lemmas 4.3 and 4.4, the crucial identity is the Lyapunov Inequality in Durrett (2019) and a couple of matrix concentration inequalities Mackey *et al.* (2014). Since there is no essential difference between Lemmas 25-27 in Zhou *et al.* (2018) and our setting, we refer readers to Zhou *et al.* (2018) and references therein.

Lemma 4.3. *Under Assumptions 1-4, for the semi-stochastic gradient \mathbf{v}_t^s and semi-stochastic Hessian \mathbf{U}_t^s , we have*

$$\mathbb{E}_{I_g} \|\text{grad}F(\mathbf{x}_t^s) - \mathbf{v}_t^s\|^{3/2} \leq \frac{L_H^{3/2}}{b_g^{3/4}} \|\text{Exp}_{\mathbf{x}_t^s}^{-1}(\mathbf{x}_t^s)\|^3, \quad (35)$$

$$\mathbb{E}_{I_h} \|\text{Hess}F(\mathbf{x}_t^s) - \mathbf{U}_t^s\|_{op}^3 \leq 64L_H^3(\rho + \rho^2)^3 \|\text{Exp}_{\mathbf{x}_t^s}^{-1}(\mathbf{x}_t^s)\|^3, \quad (36)$$

where $\rho = \sqrt{\frac{2e \log mn}{b_h}}$.

Lemma 4.4. *For any $\eta \in T_{\mathbf{x}_t^s} \mathcal{M}$ and $M > 0$, we have*

$$\langle \text{grad}F(\mathbf{x}_t^s) - \mathbf{v}_t^s, \eta \rangle \leq \frac{M}{27} \|\eta\|^3 + \frac{2\|\text{grad}F(\mathbf{x}_t^s) - \mathbf{v}_t^s\|^{3/2}}{M^{1/2}}, \quad (37)$$

$$\langle (\text{Hess}F(\mathbf{x}_t^s) - \mathbf{U}_t^s)[\eta], \eta \rangle \leq \frac{2M}{27} \|\eta\|^3 + \frac{27}{M^2} \|\text{Hess}F(\mathbf{x}_t^s) - \mathbf{U}_t^s\|_{op}^3. \quad (38)$$

The following Lemmas 4.5 and 4.6 provide an upper bound on $\|\text{grad}F\|$ and a lower bound on $\lambda_{\min}(\text{Hess}F)$, respectively.

Lemma 4.5. *Under Assumptions 1-4, if $\sigma \geq 2L_H$ and also satisfies (34), then for any $\mathbf{h} \in T_{\mathbf{x}_t^s} \mathcal{M}$ such that $\|\mathbf{h}\| < \text{inj}(\mathcal{M})$, we have*

$$\|\text{grad}F(\text{Exp}_{\mathbf{x}_t^s}(\mathbf{h}))\| \leq \sigma \|\mathbf{h}\|^2 + \|\text{grad}F(\mathbf{x}_t^s) - \mathbf{v}_t^s\| + \frac{1}{\sigma} \|\text{Hess}F(\mathbf{x}_t^s) - \mathbf{U}_t^s\|_{op}^2 + \|\nabla m_t^s(\mathbf{h})\|. \quad (39)$$

Proof. For simplicity, we denote $\text{Exp}_{\mathbf{x}_t^s}(\mathbf{h})$ by \mathbf{y} , the parallel transport operator $\Gamma_{\mathbf{x}_t^s}^{\mathbf{y}}$ by Γ , and $\Gamma_{\mathbf{y}}^{\mathbf{x}_t^s}$ by Γ^{-1} . We have

$$\begin{aligned} \|\text{grad}F(\mathbf{y})\| &= \|\Gamma^{-1} \text{grad}F(\mathbf{y})\| \\ &= \|\Gamma^{-1} \text{grad}F(\mathbf{y}) - \text{grad}F(\mathbf{x}_t^s) - \text{Hess}F(\mathbf{x}_t^s)\mathbf{h} + \mathbf{v}_t^s + \mathbf{U}_t^s\mathbf{h} + \frac{\sigma\|\mathbf{h}\|}{2}\mathbf{h} \\ &\quad + (\text{grad}F(\mathbf{x}_t^s) - \mathbf{v}_t^s) + (\text{Hess}F(\mathbf{x}_t^s) - \mathbf{U}_t^s)\mathbf{h} - \frac{\sigma\|\mathbf{h}\|}{2}\mathbf{h}\| \\ &\leq \|\Gamma^{-1} \text{grad}F(\mathbf{y}) - \text{grad}F(\mathbf{x}_t^s) - \text{Hess}F(\mathbf{x}_t^s)\mathbf{h}\| + \|\mathbf{v}_t^s + \mathbf{U}_t^s\mathbf{h} + \frac{\sigma\|\mathbf{h}\|}{2}\mathbf{h}\| \\ &\quad + \|\text{grad}F(\mathbf{x}_t^s) - \mathbf{v}_t^s\| + \|(\text{Hess}F(\mathbf{x}_t^s) - \mathbf{U}_t^s)\mathbf{h}\| + \frac{\sigma\|\mathbf{h}\|^2}{2}. \end{aligned}$$

Due to the isometric property of Γ and Lemma 3.1, we have

$$\begin{aligned} \|\Gamma^{-1} \text{grad}F(\mathbf{y}) - \text{grad}F(\mathbf{x}_t^s) - \text{Hess}F(\mathbf{x}_t^s)\mathbf{h}\| &= \|\text{grad}F(\mathbf{y}) - \Gamma \text{grad}F(\mathbf{x}_t^s) - \Gamma \text{Hess}F(\mathbf{x}_t^s)\mathbf{h}\| \\ &\leq \frac{L_H}{2} \|\mathbf{h}\|^2 \leq \frac{\sigma}{4} \|\mathbf{h}\|^2, \end{aligned}$$

where the last inequality follows from the condition $\sigma \geq 2L_H$. From the definition of $m_t^s(\cdot)$ in (27), we have $\|\mathbf{v}_t^s + \mathbf{U}_t^s \mathbf{h} + \frac{\sigma \|\mathbf{h}\|}{2} \mathbf{h}\| = \|\nabla m_t^s(\mathbf{h})\|$. Note that

$$\|(\text{Hess}F(\mathbf{x}_t^s) - \mathbf{U}_t^s) \mathbf{h}\| \leq \|\text{Hess}F(\mathbf{x}_t^s) - \mathbf{U}_t^s\|_{op} \|\mathbf{h}\| \leq \frac{1}{\sigma} \|\text{Hess}F(\mathbf{x}_t^s) - \mathbf{U}_t^s\|_{op}^2 + \frac{\sigma}{4} \|\mathbf{h}\|^2,$$

where the last inequality is due to Young's inequality. Combining these results, the proof of (39) is completed. \square

Lemma 4.6. *Under Assumptions 1-4, if $\sigma \geq 2L_H$ and also satisfies (34), then for any $\mathbf{h} \in T_{\mathbf{x}_t^s} \mathcal{M}$ such that $\|\mathbf{h}\| < \text{inj}(\mathcal{M})$, we have*

$$-\lambda_{\min}(\text{Hess}F(\text{Exp}_{\mathbf{x}_t^s}(\mathbf{h}))) \leq \sigma \|\mathbf{h}\| + \|\text{Hess}F(\mathbf{x}_t^s) - \mathbf{U}_t^s\|_{op} + \frac{\sigma}{2} \|\mathbf{h}\| - \|\mathbf{h}_t^s\|, \quad (40)$$

where $\lambda_{\min}(\text{Hess}F(\mathbf{x}))$ is defined as $\lambda_{\min}(\text{Hess}F(\mathbf{x})) = \inf_{\eta \in T_{\mathbf{x}} \mathcal{M}} \left\{ \frac{\langle \text{Hess}F(\mathbf{x}) \eta, \eta \rangle}{\|\eta\|} \right\}$.

Proof. Denote $\text{Exp}_{\mathbf{x}_t^s}(\mathbf{h})$ by \mathbf{y} and \mathbf{x}_t^s by \mathbf{x} . Furthermore, let $\mathbf{I}_{\mathbf{x}}$ denotes the identity operator at \mathbf{x} , i.e., $\mathbf{I}_{\mathbf{x}}(\eta) = \eta$ for any $\eta \in T_{\mathbf{x}} \mathcal{M}$. We have

$$\begin{aligned} H_{\mathbf{y}} &\succeq \Gamma_{\mathbf{x}}^{\mathbf{y}} H_{\mathbf{x}} \Gamma_{\mathbf{y}}^{\mathbf{x}} - L_H \|\mathbf{h}\| \mathbf{I}_{\mathbf{y}} \\ &\succeq \Gamma_{\mathbf{x}}^{\mathbf{y}} U_t^s \Gamma_{\mathbf{y}}^{\mathbf{x}} - \|\Gamma_{\mathbf{x}}^{\mathbf{y}} H_{\mathbf{x}} \Gamma_{\mathbf{y}}^{\mathbf{x}} - \Gamma_{\mathbf{x}}^{\mathbf{y}} U_t^s \Gamma_{\mathbf{y}}^{\mathbf{x}}\|_{op} \mathbf{I}_{\mathbf{y}} - L_H \|\mathbf{h}\| \mathbf{I}_{\mathbf{y}} \\ &\succeq -\frac{\sigma}{2} \|\mathbf{h}_t^s\| \mathbf{I}_{\mathbf{y}} - \|H_{\mathbf{x}} - U_t^s\|_{op} \mathbf{I}_{\mathbf{y}} - L_H \|\mathbf{h}\| \mathbf{I}_{\mathbf{y}}, \end{aligned}$$

where the first inequality follows from the H-smooth assumption (10), the second inequality follows from the definition of the operator norm and triangle inequality, and the third inequality follows from the isometric property of the parallel transport Γ and the following argument. Assume that $\Gamma_{\mathbf{x}}^{\mathbf{y}} U_t^s \Gamma_{\mathbf{y}}^{\mathbf{x}} \succeq -\frac{\sigma}{2} \|\mathbf{h}_t^s\| \mathbf{I}_{\mathbf{y}}$ does not hold, then there exists $\xi \in T_{\mathbf{y}} \mathcal{M}$, s.t. $\langle \xi, \Gamma_{\mathbf{x}}^{\mathbf{y}} U_t^s \Gamma_{\mathbf{y}}^{\mathbf{x}} \xi \rangle + \frac{\sigma}{2} \|\mathbf{h}_t^s\| \cdot \|\xi\|^2 < 0$. Denote the $\Gamma_{\mathbf{y}}^{\mathbf{x}} \xi$ by η , we have $\langle \eta, \mathbf{U}_t^s \eta \rangle + \frac{\sigma}{2} \|\mathbf{h}_t^s\| \cdot \|\eta\|^2 = \langle \xi, \Gamma_{\mathbf{x}}^{\mathbf{y}} U_t^s \Gamma_{\mathbf{y}}^{\mathbf{x}} \xi \rangle + \frac{\sigma}{2} \|\mathbf{h}_t^s\| \cdot \|\xi\|^2 < 0$, which contradicts (30). Therefore, we have

$$\begin{aligned} -\lambda_{\min}(H_{\mathbf{y}}) &\leq \frac{\sigma}{2} \|\mathbf{h}_t^s\| + \|H_{\mathbf{x}} - U_t^s\|_{op} + L_H \|\mathbf{h}\| \\ &= \frac{\sigma}{2} (\|\mathbf{h}_t^s\| - \|\mathbf{h}\|) + \|H_{\mathbf{x}} - U_t^s\|_{op} + (L_H + \sigma/2) \|\mathbf{h}\| \\ &\leq \sigma \|\mathbf{h}\| + \|H_{\mathbf{x}} - U_t^s\|_{op} + \frac{\sigma}{2} \|\mathbf{h}_t^s\| - \|\mathbf{h}\|, \end{aligned}$$

where the last inequality holds because $L_H \leq \sigma/2$. \square

Combining Lemmas 4.5 and 4.6 and the definition of $\mu(\mathbf{x})$ in (23), we have the following result.

Lemma 4.7. *Under Assumptions 1-4, setting $\sigma = \bar{k} L_H$ such that $\bar{k} \geq 2$ and σ satisfying (34), then for any $\mathbf{h} \in T_{\mathbf{x}_t^s} \mathcal{M}$ such that $\|\mathbf{h}\| < \text{inj}(\mathcal{M})$, we have*

$$\begin{aligned} \mu(\text{Exp}_{\mathbf{x}_t^s}(\mathbf{h})) &\leq 9\bar{k}^{3/2} [\sigma^{3/2} \|\mathbf{h}\|^3 + \|\text{grad}F(\mathbf{x}_t^s) - \mathbf{v}_t^s\|^{3/2} + \sigma^{-3/2} \|\text{Hess}F(\mathbf{x}_t^s) - \mathbf{U}_t^s\|^3 \\ &\quad + \|\nabla m_t^s(\mathbf{h})\|^{3/2} + \frac{\sigma^{3/2}}{8} \|\mathbf{h}\| - \|\mathbf{h}_t^s\|^3]. \end{aligned}$$

Proof. The proof follows from that of Lemma 4.14. \square

Next, We present the following result from Zhang *et al.* (2016). This inequality extends the law of cosines from Euclidean space to Riemannian space, which is fundamental to carry out non-asymptotic analysis for Riemannian optimization. The resulting inequality is used in the proof of Lemma 4.9.

Lemma 4.8 (Zhang *et al.* (2016), Lemma 5). *If a , b and c are the side lengths of a geodesic triangle in an Alexandrov space with curvature lower-bounded by κ , and A is the angle between sides b and c , then*

$$a^2 \leq \frac{\sqrt{|\kappa|}c}{\tanh \sqrt{|\kappa|}c} b^2 + c^2 - 2bc \cos A. \quad (41)$$

Lemmas 4.9 and 4.10 below are used in the proof of the first main result presented in Theorem 4.1.

Lemma 4.9. *Let $\zeta \triangleq \sqrt{|\kappa|} \text{inj}(\mathcal{M}) / \tanh \sqrt{|\kappa|} \text{inj}(\mathcal{M})$ if $\kappa < 0$ and $\zeta \triangleq 1$, o.w.. Then, under Assumptions 1-5, for any $\mathbf{h} \in T_{\mathbf{x}_t^s} \mathcal{M}$ such that $\|\mathbf{h}\| < \text{inj}(\mathcal{M})$ and $T \geq 2$, we have*

$$\| \text{Exp}_{\mathbf{x}_t^s}^{-1}(\text{Exp}_{\mathbf{x}_t^s}(\mathbf{h})) \|^3 \leq 2(\sqrt{\zeta - 1} + 1)^3 T^2 \|\mathbf{h}\|^3 + (1 + \frac{3}{T}) \|\text{Exp}_{\mathbf{x}_t^s}^{-1}(\mathbf{x}_t^s)\|^3. \quad (42)$$

Proof. Let $g(t) = t / \tanh t$ which is non-decreasing on $[0, \sqrt{|\kappa|}D]$ and $g(t) \geq 1$. For simplicity, denote $\|\text{Exp}_{\mathbf{x}_t^s}^{-1}(\text{Exp}_{\mathbf{x}_t^s}(\mathbf{h}))\|$, $\|\mathbf{h}\|$ and $\|\text{Exp}_{\mathbf{x}_t^s}^{-1}(\mathbf{x}_t^s)\|$ by a , b and c , respectively. By Lemma 4.8, we have

$$a^2 \leq \frac{\sqrt{|\kappa|}c}{\tanh \sqrt{|\kappa|}c} b^2 + c^2 - 2bc \cos A \leq (b + c)^2 + (\zeta - 1)b^2 \leq [(\sqrt{\zeta - 1} + 1)b + c]^2.$$

Therefore,

$$\begin{aligned} a^3 &\leq [(\sqrt{\zeta - 1} + 1)b + c]^3 \\ &= (\sqrt{\zeta - 1} + 1)^3 b^3 + 3T^{1/3}(\sqrt{\zeta - 1} + 1)^2 b^2 \frac{c}{T^{1/3}} + 3T^{2/3}(\sqrt{\zeta - 1} + 1)b \frac{c^2}{T^{2/3}} + c^3 \\ &\leq (\sqrt{\zeta - 1} + 1)^3 b^3 + 3(\frac{2}{3}[T^{1/3}(\sqrt{\zeta - 1} + 1)^2 b^2]^{3/2} + \frac{1}{3} \frac{c^3}{T}) + 3(\frac{1}{3}[T^{2/3}(\sqrt{\zeta - 1} + 1)b]^3 + \frac{2c^3}{3T}) + c^3 \\ &= (\sqrt{\zeta - 1} + 1)^3 (1 + 2\sqrt{T} + T^2) b^3 + (1 + \frac{3}{T}) c^3 \\ &\leq 2(\sqrt{\zeta - 1} + 1)^3 T^2 b^3 + (1 + \frac{3}{T}) c^3, \end{aligned}$$

where the second inequality follows from Young's inequality and the last inequality follows from the fact that $1 + 2\sqrt{T} + T^2 \leq 2T^2$ when $T \geq 2$. \square

Lemma 4.10. *Define the series $c_t \triangleq c_{t+1}(1 + 3/T) + \sigma[500T^3(\sqrt{\xi - 1} + 1)^3]^{-1}$ for $0 \leq t \leq T - 1$ and $c_T = 0$. Then for any $1 \leq t \leq T$, we have*

$$\sigma/24 - 2c_t(\sqrt{\xi - 1} + 1)^3 T^2 \geq 0. \quad (43)$$

Proof. Assuming $c_t + q = p(c_{t+1} + q)$, we can derive $p = 1 + 3/T$ and $q = \sigma[1500T^2(\sqrt{\xi - 1} + 1)^3]^{-1}$. Furthermore, given $c_T = 0$ by induction, we have $c_t = (p^{T-t} - 1)q$. Therefore,

$$2c_t(\sqrt{\xi - 1} + 1)^3 T^2 = ((1 + \frac{3}{T})^{T-t} - 1) \frac{\sigma}{750} \leq (1 + \frac{3}{T})^T \frac{\sigma}{750} \leq \frac{\sigma}{24}, \quad (44)$$

where the last inequality follows from the fact $(1 + 3/T)^T \leq 27$. \square

Theorem 4.1 below presents our first main result. It provides the convergence rate of the R-SVRC algorithm when the cubic regularized Newton subproblem is solved *exactly*.

Theorem 4.1. *Under Assumptions 1-5, suppose that the cubic regularization parameter σ in Algorithm 1 is fixed and satisfies $\sigma = \bar{k}L_H$, where L_H is the Hessian Lipschitz parameter according to (10), $\bar{k} \geq 2$ and σ satisfies (34). Furthermore, assume that the batch size parameters b_g and b_h satisfy*

$$b_g \geq \frac{3000^{4/3} T^4 (\sqrt{\xi - 1} + 1)^4}{\bar{k}^2}, \quad b_h \geq \frac{e \log d}{(\sqrt{\frac{\bar{k}}{193T(\sqrt{\xi - 1} + 1)}} + \frac{1}{8} - \frac{1}{2\sqrt{2}})^2}, \quad (45)$$

where $T \geq 2$ is the length of the inner loop, e is the Euler's number and $d = mn$ is the dimension of the problem. Then, we have

$$\mathbb{E}[\mu(\mathbf{x}_{out})] \leq \frac{240\bar{k}^2 L_H^{1/2} \Delta_F}{ST}, \quad (46)$$

where $\mu(\mathbf{x})$ is defined in (23).

Proof. First, we upper bound $F(\mathbf{x}_{t+1}^s)$ as follows:

$$\begin{aligned} F(\mathbf{x}_{t+1}^s) &\leq F(\mathbf{x}_t^s) + \langle \text{grad} F(\mathbf{x}_t^s), \mathbf{h}_t^s \rangle + \frac{1}{2} \langle H_{\mathbf{x}_t^s}[\mathbf{h}_t^s], \mathbf{h}_t^s \rangle + \frac{L_H}{6} \|\mathbf{h}_t^s\|^3 \\ &= F(\mathbf{x}_t^s) + \langle \text{grad} F(\mathbf{x}_t^s) - \mathbf{v}_t^s, \mathbf{h}_t^s \rangle + \frac{1}{2} \langle (H_{\mathbf{x}_t^s} - \mathbf{U}_t^s)[\mathbf{h}_t^s], \mathbf{h}_t^s \rangle - \frac{\sigma - L_H}{6} \|\mathbf{h}_t^s\|^3 \\ &\quad + \langle \mathbf{v}_t^s, \mathbf{h}_t^s \rangle + \frac{1}{2} \langle \mathbf{U}_t^s[\mathbf{h}_t^s], \mathbf{h}_t^s \rangle + \frac{\sigma}{6} \|\mathbf{h}_t^s\|^3 \\ &\leq F(\mathbf{x}_t^s) + (\frac{\sigma}{27} \|\mathbf{h}_t^s\|^3 + \frac{2}{\sigma^{1/2}} \|\text{grad} F(\mathbf{x}_t^s) - \mathbf{v}_t^s\|^{3/2}) + \frac{1}{2} (\frac{2\sigma}{27} \|\mathbf{h}_t^s\|^3 + \frac{27}{\sigma^2} \|H_{\mathbf{x}_t^s} - \mathbf{U}_t^s\|_{op}^3) \\ &\quad - \frac{\sigma - L_H}{6} \|\mathbf{h}_t^s\|^3 - \frac{\sigma}{12} \|\mathbf{h}_t^s\|^3 \\ &\leq F(\mathbf{x}_t^s) + \frac{2}{\sigma^{1/2}} \|\text{grad} F(\mathbf{x}_t^s) - \mathbf{v}_t^s\|^{3/2} + \frac{27}{2\sigma^2} \|H_{\mathbf{x}_t^s} - \mathbf{U}_t^s\|_{op}^3 - \frac{\sigma}{12} \|\mathbf{h}_t^s\|^3, \end{aligned} \quad (47)$$

where the first inequality follows from Lemma 3.1 and the second inequality holds due to Lemmas 4.4 and 4.1. Next, we define

$$R_t^s = \mathbb{E}[F(\mathbf{x}_t^s) + c_t \|\text{Exp}_{\tilde{\mathbf{x}}^s}^{-1}(\mathbf{x}_t^s)\|^3], \quad (48)$$

where c_t is defined in Lemma 4.10. By Lemma 4.9, for $T \geq 2$, we have

$$c_{t+1} \|\text{Exp}_{\tilde{\mathbf{x}}^s}^{-1}(\text{Exp}_{\mathbf{x}_t^s}(\mathbf{h}_t^s))\|^3 \leq 2c_{t+1}(\sqrt{\xi - 1} + 1)^3 T^2 \|\mathbf{h}_t^s\|^3 + c_{t+1}(1 + \frac{3}{T}) \|\text{Exp}_{\tilde{\mathbf{x}}^s}^{-1}(\mathbf{x}_t^s)\|^3. \quad (49)$$

From Lemma 4.7 with $\mathbf{h} = \mathbf{h}_t^s$ using the condition (29) and the definition of \mathbf{x}_{t+1}^s , we have

$$\frac{\mu(\mathbf{x}_{t+1}^s)}{240\bar{k}^2\sqrt{L_H}} \leq \frac{\sigma}{24}\|\mathbf{h}_t^s\|^3 + \frac{\|\text{grad}F(\mathbf{x}_t^s) - \mathbf{v}_t^s\|^{3/2}}{24\sqrt{\sigma}} + \frac{\|\text{Hess}F(\mathbf{x}_t^s) - \mathbf{U}_t^s\|^3}{24\sigma^2}. \quad (50)$$

From (47), we have

$$\begin{aligned} R_{t+1}^s &+ \mathbb{E}\left[\frac{\mu(\mathbf{x}_{t+1}^s)}{240\bar{k}^2\sqrt{L_H}}\right] \\ &= \mathbb{E}[F(\mathbf{x}_{t+1}^s) + c_{t+1}\|\text{Exp}_{\hat{\mathbf{x}}^s}^{-1}(\mathbf{x}_{t+1}^s)\|^3 + \frac{\mu(\mathbf{x}_{t+1}^s)}{240\bar{k}^2\sqrt{L_H}}] \\ &\leq \mathbb{E}[F(\mathbf{x}_t^s) + \frac{3}{\sqrt{\sigma}}\|\text{grad}F(\mathbf{x}_t^s) - \mathbf{v}_t^s\|^{3/2} + \frac{14}{\sigma^2}\|\text{Hess}F(\mathbf{x}_t^s) - \mathbf{U}_t^s\|_{op}^3] \\ &\quad + \mathbb{E}[c_{t+1}(1 + \frac{3}{T})\|\text{Exp}_{\hat{\mathbf{x}}^s}^{-1}(\mathbf{x}_t^s)\|^3 - (\frac{\sigma}{24} - 2c_{t+1}(\sqrt{\xi} - 1 + 1)^3T^2)\|\mathbf{h}_t^s\|^3] \\ &\leq \mathbb{E}[F(\mathbf{x}_t^s) + \frac{3}{\sqrt{\sigma}}\|\text{grad}F(\mathbf{x}_t^s) - \mathbf{v}_t^s\|^{3/2} + \frac{14}{\sigma^2}\|\text{Hess}F(\mathbf{x}_t^s) - \mathbf{U}_t^s\|_{op}^3 + c_{t+1}(1 + \frac{3}{T})\|\text{Exp}_{\hat{\mathbf{x}}^s}^{-1}(\mathbf{x}_t^s)\|^3], \end{aligned}$$

where the first inequality follows from (47), (49), (50) and the last inequality follows from Lemma 4.10.

Based on Lemma 4.3 and the conditions on b_g and b_h , it can be verified that

$$\begin{aligned} \frac{3}{\sqrt{\sigma}}\mathbb{E}\|\text{grad}F(\mathbf{x}_t^s) - \mathbf{v}_t^s\|^{3/2} &\leq \frac{3L_H^{3/2}}{\sqrt{\sigma}b_g^{3/4}}\mathbb{E}\|\text{Exp}_{\hat{\mathbf{x}}^s}^{-1}(\mathbf{x}_t^s)\|^3 \leq \frac{\sigma}{1000T^3(\sqrt{\zeta} - 1 + 1)^3}\mathbb{E}\|\text{Exp}_{\hat{\mathbf{x}}^s}^{-1}(\mathbf{x}_t^s)\|^3, \\ \frac{14}{\sigma^2}\mathbb{E}\|\text{Hess}F(\mathbf{x}_t^s) - \mathbf{U}_t^s\|_{op}^3 &\leq \frac{896L_H^3(\rho + \rho^2)^3}{\sigma^2}\mathbb{E}\|\text{Exp}_{\hat{\mathbf{x}}^s}^{-1}(\mathbf{x}_t^s)\|^3 \leq \frac{\sigma}{1000T^3(\sqrt{\zeta} - 1 + 1)^3}\mathbb{E}\|\text{Exp}_{\hat{\mathbf{x}}^s}^{-1}(\mathbf{x}_t^s)\|^3, \end{aligned}$$

where $\rho = \sqrt{\frac{2e \log mn}{b_h}}$. Therefore, we have

$$\begin{aligned} R_{t+1}^s &+ \mathbb{E}\left[\frac{\mu(\mathbf{x}_{t+1}^s)}{240\bar{k}^2\sqrt{L_H}}\right] \leq \mathbb{E}[F(\mathbf{x}_t^s) + \|\text{Exp}_{\hat{\mathbf{x}}^s}^{-1}(\mathbf{x}_t^s)\|^3(c_{t+1}(1 + 3/T) + \frac{\sigma}{500T^3(\sqrt{\zeta} - 1 + 1)^3})] \\ &= \mathbb{E}[F(\mathbf{x}_t^s) + c_t\|\text{Exp}_{\hat{\mathbf{x}}^s}^{-1}(\mathbf{x}_t^s)\|^3] = R_t^s, \end{aligned}$$

where the first equality comes from the definition of c_t in Lemma 4.10. Telescoping the above inequality from $t = 0$ to $T - 1$, we have

$$R_0^s - R_T^s \geq (240\bar{k}^2\sqrt{L_H})^{-1} \sum_{t=1}^T \mathbb{E}[\mu(\mathbf{x}_t^s)].$$

Note that $c_T = 0$ and $\mathbf{x}_T^{s-1} = \mathbf{x}_0^s = \hat{\mathbf{x}}^s$, then $R_T^s = \mathbb{E}[F(\mathbf{x}_T^s) + c_T\|\text{Exp}_{\hat{\mathbf{x}}^s}^{-1}(\mathbf{x}_T^s)\|^3] = \mathbb{E}F(\hat{\mathbf{x}}^{s+1})$ and $R_0^s = \mathbb{E}[F(\mathbf{x}_0^s) + c_0\|\text{Exp}_{\hat{\mathbf{x}}^s}^{-1}(\mathbf{x}_0^s)\|^3] = \mathbb{E}F(\hat{\mathbf{x}}^s)$, which implies

$$\mathbb{E}F(\hat{\mathbf{x}}^s) - \mathbb{E}F(\hat{\mathbf{x}}^{s+1}) = R_0^s - R_T^s \geq (240\bar{k}^2\sqrt{L_H})^{-1} \sum_{t=1}^T \mathbb{E}[\mu(\mathbf{x}_t^s)].$$

Telescoping the above inequality from $s = 1$ to S yields

$$\Delta_F \geq \sum_{s=1}^S \mathbb{E}F(\hat{\mathbf{x}}^s) - \mathbb{E}F(\hat{\mathbf{x}}^{s+1}) \geq (240\bar{k}^2\sqrt{L_H})^{-1} \sum_{s=1}^S \sum_{t=1}^T \mathbb{E}[\mu(\mathbf{x}_t^s)].$$

By the definition of the choice of \mathbf{x}_{out} , the proof is completed. \square

Remark 7. Let $K = ST$ where S and T are the number of epochs and epoch length in Algorithm 1. Following our discussion below (24) and by Theorem 4.1, setting $\mathbb{E}[\mu(\mathbf{x})] \leq 240\bar{k}^2 L_H^{1/2} \Delta_F / K \leq \epsilon^{3/2}$, the algorithm obtains a $(\epsilon, \sqrt{\epsilon})$ -solution in $O(\epsilon^{-3/2})$ iterations. In other words, the algorithm obtains a first-order stationary point (i.e., $\|\text{grad}F(\mathbf{x})\| \leq \epsilon$) in $O(\epsilon^{-3/2})$ iterations and a second-order stationary point (i.e., $\lambda_{\min}(\text{Hess}F(\mathbf{x})) \geq -\epsilon$) in $O(\epsilon^{-3})$ iterations.

Definition 4.3 (Second-order oracle). Given an index i and a point \mathbf{x} , a second-order oracle (SO) call returns a triple $[f_i(\mathbf{x}), \nabla f_i(\mathbf{x}), \nabla^2 f_i(\mathbf{x})]$.

When manifold is embedded in a Euclidean space, calculating the Riemannian gradient and Hessian (applied to a certain direction) requires the Euclidean gradient and Hessian. Therefore, the number of SO calls is a reasonable metric to evaluate complexities of different algorithms, stochastic and deterministic. In numerical studies, we also compare different methods on the number of SO calls.

Corollary 4.1. Suppose that the cubic regularization parameter σ in Algorithm 1 is fixed and satisfies $\sigma = \bar{k}L_H$, where L_H is the Hessian Lipschitz parameter according to (10), $\bar{k} \geq 2$ and σ satisfies (34). Let the epoch length $T = N^{1/5}$, batch sizes $b_g = \frac{3000^{4/3} N^{4/5} (\sqrt{\zeta-1}+1)^4}{\bar{k}^2}$, $b_h = \frac{e \log d}{(\sqrt{\frac{\bar{k}}{193N^{1/5}(\sqrt{\zeta-1}+1)} + \frac{1}{8} - \frac{1}{2\sqrt{2}}})^2}$, and the number of epochs $S = \max\{1, 240\bar{k}^2 L_H^{1/2} \Delta_F N^{-1/5} \epsilon^{-3/2}\}$, where $d = mn$ is the dimension of the problem. Then, under Assumptions 1-5, Algorithm 1 finds an $(\epsilon, \sqrt{L_H \epsilon})$ -second-order stationary point in $\tilde{O}(N + L_H^{1/2} \Delta_F N^{4/5} \epsilon^{-3/2})$ second-order oracle calls.

Proof. The parameter setting in Corollary 4.1 satisfies the requirements of Theorem 4.1. The epoch size S enforce $\mathbb{E}[\mu(\mathbf{x}_{out})] \leq \epsilon$, which implies that \mathbf{x}_{out} is an $(\epsilon, \sqrt{L_H \epsilon})$ -approximate local minimum. Note that Algorithm 1 requires calculating full gradient ∇F and Hessian $\nabla^2 F$ at the beginning of each epoch with N SO calls. Inside each epoch, it needs to calculate stochastic gradient and Hessian with $b_g + b_h$ SO calls at each iteration. Thus, the total number of SO calls is

$$\begin{aligned} SN + (ST)(b_g + b_h) &\leq N + 240\bar{k}^2 L_H^{1/2} \Delta_F N^{4/5} \epsilon^{-3/2} + 240\bar{k}^2 L_H^{1/2} \Delta_F \epsilon^{-3/2} (b_g + b_h) \\ &= \tilde{O}(N + L_H^{1/2} \Delta_F N^{4/5} \epsilon^{-3/2}), \end{aligned}$$

where the \tilde{O} comes from $\log d$ in b_h . \square

In practice, finding the exact solution to the cubic-regularized Newton subproblem (27) is not always computationally desirable Agarwal *et al.* (2020), Nesterov & Polyak (2006), Cartis *et al.* (2011a,b). Instead, we can solve the subproblem *inexactly*, but yet guarantee theoretical properties of the algorithm. More specifically, we propose to solve the cubic-regularized Newton subproblem *inexactly*, but the one that satisfies the conditions in Definition 4.4 below. It is then proved in Theorem 4.2 that the complexity of the algorithm with *inexact* solution to its subproblem is the same as the original algorithm, except for an $O(1)$ constant.

Definition 4.4 (Inexact solution). *Given a $\delta > 0$, $\tilde{\mathbf{h}}_t^s$ is a δ -inexact solution to (27) if it satisfies*

$$m_t^s(\tilde{\mathbf{h}}_t^s) \leq -\frac{\sigma}{12} \|\tilde{\mathbf{h}}_t^s\|^3 + \delta, \quad (51)$$

$$\|\nabla m_t^s(\tilde{\mathbf{h}}_t^s)\| \leq (\sigma)^{1/3} \delta^{2/3}, \quad (52)$$

$$\lambda_{\min}(\nabla^2 m_t^s(\tilde{\mathbf{h}}_t^s)) \geq -(\sigma)^{2/3} \delta^{1/3}. \quad (53)$$

The following lemma is parallel to Lemma 4.1 when the subproblem is solved inexactly.

Lemma 4.11. *Under Assumption 3, if $\tilde{\mathbf{h}}_t^s$ is a δ -inexact solution to (27), then*

$$\langle \mathbf{v}_t^s, \tilde{\mathbf{h}}_t^s \rangle + \frac{1}{2} \langle \mathbf{U}_t^s \tilde{\mathbf{h}}_t^s, \tilde{\mathbf{h}}_t^s \rangle + \frac{\sigma}{6} \|\tilde{\mathbf{h}}_t^s\|^3 \leq -\frac{\sigma}{12} \|\tilde{\mathbf{h}}_t^s\|^3 + \delta, \quad (54)$$

$$\|\mathbf{v}_t^s + \mathbf{U}_t^s \tilde{\mathbf{h}}_t^s + (\frac{\sigma}{2} \|\tilde{\mathbf{h}}_t^s\|) \tilde{\mathbf{h}}_t^s\| \leq (\sigma)^{1/3} \delta^{2/3}, \quad (55)$$

$$\mathbf{U}_t^s + \sigma \|\tilde{\mathbf{h}}_t^s\| \mathbf{I} \succeq -(\sigma)^{2/3} \delta^{1/3} \mathbf{I}. \quad (56)$$

Proof. Inequalities (54) and (55) follow from expanding $m_t^s(\tilde{\mathbf{h}}_t^s)$ and $\nabla m_t^s(\tilde{\mathbf{h}}_t^s)$ in (51) and (52). To show (56), note that $\nabla^2 m_t^s(\tilde{\mathbf{h}}_t^s) = \mathbf{U}_t^s + \lambda \mathbf{I} + \lambda (\frac{\tilde{\mathbf{h}}_t^s}{\|\tilde{\mathbf{h}}_t^s\|}) (\frac{\tilde{\mathbf{h}}_t^s}{\|\tilde{\mathbf{h}}_t^s\|})^\top$, where $\lambda = \frac{\sigma \|\tilde{\mathbf{h}}_t^s\|}{2}$. We have

$$\mathbf{U}_t^s + 2\lambda \mathbf{I} \succeq \mathbf{U}_t^s + \lambda \mathbf{I} + \lambda (\frac{\tilde{\mathbf{h}}_t^s}{\|\tilde{\mathbf{h}}_t^s\|}) (\frac{\tilde{\mathbf{h}}_t^s}{\|\tilde{\mathbf{h}}_t^s\|})^\top \succeq -(\sigma)^{2/3} \delta^{1/3} \mathbf{I},$$

where the first inequality follows from the Cauchy–Schwarz inequality, $\|\mathbf{v}\| \geq \frac{\langle \mathbf{v}, \tilde{\mathbf{h}}_t^s \rangle}{\|\tilde{\mathbf{h}}_t^s\|}$ for any $\mathbf{v} \in \mathbb{R}^{m \times n}$, and the second inequality follows from (56). \square

Parallel to Lemma 4.2, Lemma 4.12 provides an upper bound on $\|\tilde{\mathbf{h}}_t^s\|$ which then provides a required (lower) bound on the cubic regularization parameter σ to have the iterates close enough to the epoch points - see Remark 8.

Lemma 4.12. *Under Assumption 1-4, given a constant $C > 0$ and $\sigma > [\frac{\delta^{2/3} + \sqrt{\delta^{4/3} + 2C^2 \cdot (C \cdot c_H + c_g)}}{C^2}]^2$, we have $\|\tilde{\mathbf{h}}_t^s\| < C$.*

Proof. Based on (55) and Cauchy–Schwarz inequality, we have

$$\langle \mathbf{v}_t^s, \tilde{\mathbf{h}}_t^s \rangle + \langle \mathbf{U}_t^s \tilde{\mathbf{h}}_t^s, \tilde{\mathbf{h}}_t^s \rangle + \frac{\sigma}{2} \|\tilde{\mathbf{h}}_t^s\|^3 \leq (\sigma)^{1/3} \delta^{2/3} \cdot \|\tilde{\mathbf{h}}_t^s\|. \quad (57)$$

which implies,

$$\frac{\sigma}{2} \|\tilde{\mathbf{h}}_t^s\|^3 \leq \|\mathbf{v}_t^s\| \cdot \|\tilde{\mathbf{h}}_t^s\| + \|\mathbf{U}_t^s\|_{op} \cdot \|\tilde{\mathbf{h}}_t^s\|^2 + (\sigma)^{1/3} \delta^{2/3} \cdot \|\tilde{\mathbf{h}}_t^s\|. \quad (58)$$

Based on (25) and dividing both sides by $\|\tilde{\mathbf{h}}_t^s\|$, we have

$$\frac{\sigma}{2} \|\tilde{\mathbf{h}}_t^s\|^2 - c_H \cdot \|\tilde{\mathbf{h}}_t^s\| - (+\sigma^{1/3} \delta^{2/3}) \leq 0, \quad (59)$$

which implies

$$\|\tilde{\mathbf{h}}_t^s\| \leq \frac{c_H + \sqrt{c_H^2 + 2\sigma(c_g + \sigma^{1/3} \delta^{2/3})}}{\sigma} \leq \frac{c_H + \sqrt{c_H^2 + 2\sigma(c_g + \sigma^{1/2} \delta^{2/3})}}{\sigma}. \quad (60)$$

Note that the right hand side of (60) is a monotonic decreasing function on σ . After some simple manipulation, we derive that if $\sigma > [\frac{\delta^{2/3} + \sqrt{\delta^{4/3} + 2C^{2 \cdot (C \cdot c_H + c_g)}}}{C^2}]^2$, then the right hand side of (60) is upper bounded by C , which implies $\|\tilde{\mathbf{h}}_t^s\| < C$. \square

Remark 8. Using Lemma 4.2, setting $C = \text{inj}(\mathcal{M})/T$, where T is the epoch length of the algorithm, we have

$$\sigma > \left[\frac{T^2 \delta^{2/3} + \sqrt{T^4 \delta^{4/3} + 2(\text{inj}(\mathcal{M}))^2 (\text{inj}(\mathcal{M}) T c_H + c_g T^2)}}{(\text{inj}(\mathcal{M}))^2} \right]^2. \quad (61)$$

Given the lower bound on σ , for any epoch s and any iteration t inside this epoch, we have

$$d(\hat{\mathbf{x}}^s, \mathbf{x}_t^s) \leq \sum_{i=1}^t d(\mathbf{x}_{i-1}^s, \mathbf{x}_i^s) \leq \sum_{i=1}^t \|\mathbf{h}_i^s\| < \text{inj}(\mathcal{M}). \quad (62)$$

Since it is difficult to quantify the difference of \mathbf{h} with the exact solution \mathbf{h}_t^s , i.e. $\|\mathbf{h}\| - \|\mathbf{h}_t^s\|$, we need to establish results similar to Lemmas 4.6 and 4.7 based on (56).

Lemma 4.13. Let $\tilde{\mathbf{h}}_t^s$ be a δ -inexact solution to (27) with $\sigma \geq 2L_H$ that satisfies (61), then under Assumptions 1-4, for any $\mathbf{h} \in \mathbb{R}^{m \times n}$ such that $\|\mathbf{h}\| < \text{inj}(\mathcal{M})$, we have

$$-\lambda_{\min}(\text{Hess}F(\text{Exp}_{\mathbf{x}_t^s}(\mathbf{h}))) \leq \frac{3\sigma}{2} \|\mathbf{h}\| + \|H_{\mathbf{x}} - U_t^s\|_{op} + \sigma \|\tilde{\mathbf{h}}_t^s\| - \|\mathbf{h}\| + (\sigma)^{2/3} \delta^{1/3}, \quad (63)$$

where $\lambda_{\min}(\text{Hess}F(\mathbf{x}))$ is defined as $\lambda_{\min}(\text{Hess}F(\mathbf{x})) \triangleq \inf_{\eta \in T_{\mathbf{x}}\mathcal{M}} \left\{ \frac{\langle \text{Hess}F(\mathbf{x})\eta, \eta \rangle}{\|\eta\|} \right\}$.

Proof. Denote $\text{Exp}_{\mathbf{x}_t^s}(\mathbf{h})$ by \mathbf{y} and \mathbf{x}_t^s by \mathbf{x} . Furthermore, let the identity operator at \mathbf{x} be denoted by $\mathbf{I}_{\mathbf{x}}$, i.e. $\mathbf{I}_{\mathbf{x}}(\eta) = \eta$ for any $\eta \in T_{\mathbf{x}}\mathcal{M}$. We have

$$\begin{aligned} H_{\mathbf{y}} &\succeq \Gamma_{\mathbf{x}}^{\mathbf{y}} H_{\mathbf{x}} \Gamma_{\mathbf{y}}^{\mathbf{x}} - L_H \|\mathbf{h}\| \mathbf{I}_{\mathbf{y}} \\ &\succeq \Gamma_{\mathbf{x}}^{\mathbf{y}} U_t^s \Gamma_{\mathbf{y}}^{\mathbf{x}} - \|\Gamma_{\mathbf{x}}^{\mathbf{y}} H_{\mathbf{x}} \Gamma_{\mathbf{y}}^{\mathbf{x}} - \Gamma_{\mathbf{x}}^{\mathbf{y}} U_t^s \Gamma_{\mathbf{y}}^{\mathbf{x}}\|_{op} \mathbf{I}_{\mathbf{y}} - L_H \|\mathbf{h}\| \mathbf{I}_{\mathbf{y}} \\ &\succeq -(\sigma \|\tilde{\mathbf{h}}_t^s\| + (\sigma)^{2/3} \delta^{1/3}) \mathbf{I}_{\mathbf{y}} - \|H_{\mathbf{x}} - U_t^s\|_{op} \mathbf{I}_{\mathbf{y}} - L_H \|\mathbf{h}\| \mathbf{I}_{\mathbf{y}}, \end{aligned}$$

where the first inequality follows from (10), the second inequality follows from the definition of the operator norm and the triangle inequality, and the third inequality follows from the isometric property of the parallel transport Γ and (56). Therefore, we have

$$\begin{aligned} -\lambda_{\min}(H_{\mathbf{y}}) &\leq (\sigma \|\tilde{\mathbf{h}}_t^s\| + (\sigma)^{2/3} \delta^{1/3}) + \|H_{\mathbf{x}} - U_t^s\|_{op} + L_H \|\mathbf{h}\| \\ &= \sigma (\|\tilde{\mathbf{h}}_t^s\| - \|\mathbf{h}\|) + \|H_{\mathbf{x}} - U_t^s\|_{op} + (L_H + \sigma) \|\mathbf{h}\| + (\sigma)^{2/3} \delta^{1/3} \\ &\leq \frac{3\sigma}{2} \|\mathbf{h}\| + \|H_{\mathbf{x}} - U_t^s\|_{op} + \sigma \|\tilde{\mathbf{h}}_t^s\| - \|\mathbf{h}\| + (\sigma)^{2/3} \delta^{1/3}, \end{aligned}$$

where the last inequality holds because $L_H \leq \sigma/2$. \square

Recall the $\mu(\mathbf{x})$ definition in (23), combining Lemmas 4.5 and 4.13, we have the following result.

Lemma 4.14. Setting $\sigma = \bar{k}L_H$ with $\bar{k} \geq 2$ such that it satisfies (61), under Assumptions 1-4, for any δ -inexact solution $\tilde{\mathbf{h}}_t^s$, we have

$$\mu(\text{Exp}_{\mathbf{x}_t^s}(\tilde{\mathbf{h}}_t^s)) \leq 9(\bar{k})^{3/2} \left[\frac{27(\sigma)^{3/2}}{8} \|\tilde{\mathbf{h}}_t^s\|^3 + \|\text{grad}F(\mathbf{x}_t^s) - \mathbf{v}_t^s\|^3 + (\sigma)^{-3/2} \|\text{Hess}F(\mathbf{x}_t^s) - \mathbf{U}_t^s\|_{op}^3 + (\sigma)^{1/2} \delta \right].$$

Proof. Recall $\mu(\mathbf{x}) = \max\{\|\text{grad}F(\mathbf{x})\|^{3/2}, -L_H^{-3/2}\lambda_{\min}^3(\text{Hess}F(\mathbf{x}))\}$. Next, we apply Lemmas 4.5 and 4.13 to upper bound $\|\text{grad}F(\mathbf{x})\|^{3/2}$ and $-(L_H^{3/2})^{-1}[\lambda_{\min}(\text{Hess}F(\mathbf{x}))]^3$, respectively.

$$\begin{aligned}
& \|\text{grad}F(\text{Exp}_{\mathbf{x}_t^s}(\tilde{\mathbf{h}}_t^s))\|^{3/2} \\
& \leq [\sigma\|\tilde{\mathbf{h}}_t^s\|^2 + \|\text{grad}F(\mathbf{x}_t^s) - \mathbf{v}_t^s\| + \frac{1}{\sigma}\|\text{Hess}F(\mathbf{x}_t^s) - \mathbf{U}_t^s\|_{op}^2 + \|\nabla m_t^s(\tilde{\mathbf{h}}_t^s)\|]^{3/2} \\
& \leq 2[(\sigma)^{3/2}\|\tilde{\mathbf{h}}_t^s\|^3 + \|\text{grad}F(\mathbf{x}_t^s) - \mathbf{v}_t^s\|^{3/2} + (\sigma)^{-3/2}\|\text{Hess}F(\mathbf{x}_t^s) - \mathbf{U}_t^s\|_{op}^3 + \|\nabla m_t^s(\tilde{\mathbf{h}}_t^s)\|^{3/2}] \\
& \leq 2[(\sigma)^{3/2}\|\tilde{\mathbf{h}}_t^s\|^3 + \|\text{grad}F(\mathbf{x}_t^s) - \mathbf{v}_t^s\|^{3/2} + (\sigma)^{-3/2}\|\text{Hess}F(\mathbf{x}_t^s) - \mathbf{U}_t^s\|_{op}^3 + (\sigma)^{1/2}\delta],
\end{aligned}$$

where the first inequality follows from Lemma 4.5, the second inequality holds due to the inequality $(a + b + c + d)^{3/2} \leq 2(a^{3/2} + b^{3/2} + c^{3/2} + d^{3/2})$, and the third inequality follows from (52).

$$\begin{aligned}
-L_H^{-3/2}[\lambda_{\min}(\text{Hess}F(\tilde{\mathbf{h}}_t^s))]^3 &= -(\bar{k})^{3/2}(\sigma)^{-3/2}[\lambda_{\min}(\text{Hess}F(\tilde{\mathbf{h}}_t^s))]^3 \\
&\leq (\bar{k})^{3/2}(\sigma)^{-3/2}[\frac{3\sigma}{2}\|\tilde{\mathbf{h}}_t^s\| + \|\text{Hess}F(\mathbf{x}_t^s) - \mathbf{U}_t^s\|_{op} + (\sigma)^{2/3}\delta^{1/3}]^3 \\
&\leq 9(\bar{k})^{3/2}[\frac{27(\sigma)^{3/2}}{8}\|\tilde{\mathbf{h}}_t^s\|^3 + (\sigma)^{-3/2}\|\text{Hess}F(\mathbf{x}_t^s) - \mathbf{U}_t^s\|_{op}^3 + (\sigma)^{1/2}\delta],
\end{aligned}$$

where the equality follows from $\sigma = \bar{k}L_H$, the first inequality follows from Lemma 4.13, and the last inequality follows from the inequality $(a + b + c)^3 \leq 9(a^3 + b^3 + c^3)$. Since $9(\bar{k})^{3/2} > 2$, we have

$$\begin{aligned}
\mu(\text{Exp}_{\mathbf{x}_t^s}(\tilde{\mathbf{h}}_t^s)) &= \max\{\|\text{grad}F(\text{Exp}_{\mathbf{x}_t^s}(\tilde{\mathbf{h}}_t^s))\|^{3/2}, -L_H^{-3/2}\lambda_{\min}^3(\text{Hess}F(\text{Exp}_{\mathbf{x}_t^s}(\tilde{\mathbf{h}}_t^s)))\} \\
&\leq 9(\bar{k})^{3/2}[\frac{27(\sigma)^{3/2}}{8}\|\tilde{\mathbf{h}}_t^s\|^3 + \|\text{grad}F(\mathbf{x}_t^s) - \mathbf{v}_t^s\|^{3/2} + (\sigma)^{-3/2}\|\text{Hess}F(\mathbf{x}_t^s) - \mathbf{U}_t^s\|_{op}^3 + (\sigma)^{1/2}\delta],
\end{aligned}$$

which completes the proof. \square

Theorem 4.2 below provides the convergence rate of the R-SVRC algorithm when the cubic regularized Newton subproblem is solved *inexactly*.

Theorem 4.2. *Suppose that the cubic regularization parameter σ in Algorithm 1 is fixed and satisfies $\sigma = \bar{k}L_H$, where L_H is the Hessian Lipschitz parameter according to (10) and $\bar{k} \geq 2$ and it also satisfies (61). At each iteration, let the cubic subproblem (27) be solved *inexactly* so that the results $\{\tilde{\mathbf{h}}_t^s\}$ are δ -inexact solutions. Furthermore, suppose that the batch sizes b_g and b_h satisfy*

$$b_g \geq \frac{3000^{4/3}T^4(\sqrt{\xi-1}+1)^4}{\bar{k}^2}, \quad b_h \geq \frac{e \log d}{(\sqrt{\frac{\bar{k}}{193T(\sqrt{\xi-1}+1)}} + \frac{1}{8} - \frac{1}{2\sqrt{2}})^2}, \quad (64)$$

where $T \geq 2$ is the length of the inner loop of the algorithm and $d = mn$ is the dimension of the problem. Then, under Assumptions 1-5, the output of the algorithm satisfies

$$\mathbb{E}[\mu(\mathbf{x}_{out})] \leq \frac{729\bar{k}^2L_H^{1/2}\Delta_F}{ST} + 738\bar{k}^2\sqrt{L_H}\delta, \quad (65)$$

where $\mu(\mathbf{x})$ is defined in (23).

Proof. First, we upper bound $F(\mathbf{x}_{t+1}^s)$ as follows:

$$\begin{aligned}
F(\mathbf{x}_{t+1}^s) &\leq F(\mathbf{x}_t^s) + \left\langle \text{grad}F(\mathbf{x}_t^s), \tilde{\mathbf{h}}_t^s \right\rangle + \frac{1}{2} \left\langle H_{\mathbf{x}_t^s}[\tilde{\mathbf{h}}_t^s], \tilde{\mathbf{h}}_t^s \right\rangle + \frac{L_H}{6} \|\tilde{\mathbf{h}}_t^s\|^3 \\
&= F(\mathbf{x}_t^s) + \left\langle \text{grad}F(\mathbf{x}_t^s) - \mathbf{v}_t^s, \tilde{\mathbf{h}}_t^s \right\rangle + \frac{1}{2} \left\langle (H_{\mathbf{x}_t^s} - \mathbf{U}_t^s)[\tilde{\mathbf{h}}_t^s], \tilde{\mathbf{h}}_t^s \right\rangle - \frac{\sigma - L_H}{6} \|\tilde{\mathbf{h}}_t^s\|^3 \\
&\quad + \left\langle \mathbf{v}_t^s, \tilde{\mathbf{h}}_t^s \right\rangle + \frac{1}{2} \left\langle \mathbf{U}_t^s[\tilde{\mathbf{h}}_t^s], \tilde{\mathbf{h}}_t^s \right\rangle + \frac{\sigma}{6} \|\tilde{\mathbf{h}}_t^s\|^3 \\
&\leq F(\mathbf{x}_t^s) + \left(\frac{\sigma}{27} \|\tilde{\mathbf{h}}_t^s\|^3 + \frac{2}{\sigma^{1/2}} \|\text{grad}F(\mathbf{x}_t^s) - \mathbf{v}_t^s\|^{3/2} \right) + \frac{1}{2} \left(\frac{2\sigma}{27} \|\tilde{\mathbf{h}}_t^s\|^3 + \frac{27}{\sigma^2} \|H_{\mathbf{x}_t^s} - \mathbf{U}_t^s\|_{op}^3 \right) \\
&\quad - \frac{\sigma - L_H}{6} \|\tilde{\mathbf{h}}_t^s\|^3 - \frac{\sigma}{12} \|\tilde{\mathbf{h}}_t^s\|^3 + \delta \\
&\leq F(\mathbf{x}_t^s) + \frac{2}{\sigma^{1/2}} \|\text{grad}F(\mathbf{x}_t^s) - \mathbf{v}_t^s\|^{3/2} + \frac{27}{2\sigma^2} \|H_{\mathbf{x}_t^s} - \mathbf{U}_t^s\|_{op}^3 - \frac{\sigma}{12} \|\tilde{\mathbf{h}}_t^s\|^3 + \delta,
\end{aligned}$$

where the first inequality follows from H-smooth assumption and Lemma 3.1, and the second inequality holds due to Lemma 4.4 and (54) in Lemma 4.11.

Next, we define

$$R_t^s \triangleq \mathbb{E}[F(\mathbf{x}_t^s) + c_t \|\text{Exp}_{\tilde{\mathbf{x}}^s}^{-1}(\mathbf{x}_t^s)\|^3], \quad (66)$$

where c_t is defined in Lemma 4.10. By Lemma 4.9, for $T \geq 2$, we have

$$c_{t+1} \|\text{Exp}_{\tilde{\mathbf{x}}^s}^{-1}(\text{Exp}_{\mathbf{x}_t^s}(\tilde{\mathbf{h}}_t^s))\|^3 \leq 2c_{t+1}(\sqrt{\xi-1}+1)^3 T^2 \|\tilde{\mathbf{h}}_t^s\|^3 + c_{t+1}(1 + \frac{3}{T}) \|\text{Exp}_{\tilde{\mathbf{x}}^s}^{-1}(\mathbf{x}_t^s)\|^3. \quad (67)$$

Furthermore, from Lemma 4.14, we have

$$\frac{\mu(\mathbf{x}_{t+1}^s)}{729\bar{k}^2\sqrt{L_H}} \leq \frac{\sigma}{24} \|\tilde{\mathbf{h}}_t^s\|^3 + \frac{\|\text{grad}F(\mathbf{x}_t^s) - \mathbf{v}_t^s\|^{3/2}}{81\sqrt{\sigma}} + \frac{\|\text{Hess}F(\mathbf{x}_t^s) - \mathbf{U}_t^s\|^3}{81\sigma^2} + \frac{\delta}{81}. \quad (68)$$

Combining (66), (67) and (68), we have

$$\begin{aligned}
R_{t+1}^s &+ \mathbb{E}\left[\frac{\mu(\mathbf{x}_{t+1}^s)}{729\bar{k}^2\sqrt{L_H}}\right] \\
&= \mathbb{E}[F(\mathbf{x}_{t+1}^s) + c_{t+1} \|\text{Exp}_{\tilde{\mathbf{x}}^s}^{-1}(\mathbf{x}_{t+1}^s)\|^3 + \frac{\mu(\mathbf{x}_{t+1}^s)}{729\bar{k}^2\sqrt{L_H}}] \\
&\leq \mathbb{E}[F(\mathbf{x}_t^s) + \frac{3}{\sqrt{\sigma}} \|\text{grad}F(\mathbf{x}_t^s) - \mathbf{v}_t^s\|^{3/2} + \frac{14}{\sigma^2} \|\text{Hess}F(\mathbf{x}_t^s) - \mathbf{U}_t^s\|^3] \\
&\quad + \mathbb{E}[c_{t+1}(1 + \frac{3}{T}) \|\text{Exp}_{\tilde{\mathbf{x}}^s}^{-1}(\mathbf{x}_t^s)\|^3 - (\frac{\sigma}{24} - 2c_{t+1}(\sqrt{\xi-1}+1)^3 T^2) \|\tilde{\mathbf{h}}_t^s\|^3] + \frac{82\delta}{81} \\
&\leq \mathbb{E}[F(\mathbf{x}_t^s) + \frac{3}{\sqrt{\sigma}} \|\text{grad}F(\mathbf{x}_t^s) - \mathbf{v}_t^s\|^{3/2} + \frac{14}{\sigma^2} \|\text{Hess}F(\mathbf{x}_t^s) - \mathbf{U}_t^s\|_{op}^3 + c_{t+1}(1 + \frac{3}{T}) \|\text{Exp}_{\tilde{\mathbf{x}}^s}^{-1}(\mathbf{x}_t^s)\|^3] + \frac{82\delta}{81},
\end{aligned}$$

where the last inequality follows from Lemma 4.10.

Based on Lemma 4.3 and the conditions on the sizes of b_g and b_h , we have

$$\frac{3}{\sqrt{\sigma}} \mathbb{E} \|\text{grad} F(\mathbf{x}_t^s) - \mathbf{v}_t^s\|^{3/2} \leq \frac{3L_H^{3/2}}{\sqrt{\sigma}b_g^{3/4}} \mathbb{E} \|\text{Exp}_{\hat{\mathbf{x}}^s}^{-1}(\mathbf{x}_t^s)\|^3 \leq \frac{\sigma}{1000T^3(\sqrt{\xi-1}+1)^3} \mathbb{E} \|\text{Exp}_{\hat{\mathbf{x}}^s}^{-1}(\mathbf{x}_t^s)\|^3, \quad (69)$$

$$\frac{14}{\sigma^2} \mathbb{E} \|\text{Hess} F(\mathbf{x}_t^s) - \mathbf{U}_t^s\|^3 \leq \frac{896L_H^3(\rho + \rho^2)^3}{\sigma^2} \mathbb{E} \|\text{Exp}_{\hat{\mathbf{x}}^s}^{-1}(\mathbf{x}_t^s)\|^3 \leq \frac{\sigma}{1000T^3(\sqrt{\xi-1}+1)^3} \mathbb{E} \|\text{Exp}_{\hat{\mathbf{x}}^s}^{-1}(\mathbf{x}_t^s)\|^3, \quad (70)$$

where $\rho = \sqrt{\frac{2e \log mn}{b_h}}$. Therefore, we have

$$\begin{aligned} R_{t+1}^s + \mathbb{E} \left[\frac{\mu(\mathbf{x}_{t+1}^s)}{729\bar{k}^2\sqrt{L_H}} \right] &\leq \mathbb{E} [F(\mathbf{x}_t^s) + \|\text{Exp}_{\hat{\mathbf{x}}^s}^{-1}(\mathbf{x}_t^s)\|^3 (c_{t+1}(1+3/T) + \frac{\sigma}{500T^3(\sqrt{\xi-1}+1)^3})] + \frac{82\delta}{81} \\ &= \mathbb{E} [F(\mathbf{x}_t^s) + c_t \|\text{Exp}_{\hat{\mathbf{x}}^s}^{-1}(\mathbf{x}_t^s)\|^3] + \frac{82\delta}{81} = R_t^s + \frac{82\delta}{81}, \end{aligned}$$

where the first equality is due to the choice of $\{c_t\}$ defined in Lemma 4.10. Telescoping the above inequality from $t = 0$ to $T - 1$, we have $R_0^s - R_T^s \geq (729\bar{k}^2\sqrt{L_H})^{-1} \sum_{t=1}^T (\mathbb{E}[\mu(\mathbf{x}_t^s)] - \frac{82\delta}{81})$. Note that $c_T = 0$ and $\mathbf{x}_T^{s-1} = \mathbf{x}_0^s = \hat{\mathbf{x}}^s$, then $R_T^s = \mathbb{E}[F(\mathbf{x}_T^s) + c_T \|\text{Exp}_{\hat{\mathbf{x}}^s}^{-1}(\mathbf{x}_T^s)\|^3] = \mathbb{E}F(\hat{\mathbf{x}}^{s+1})$ and $R_0^s = \mathbb{E}[F(\mathbf{x}_0^s) + c_0 \|\text{Exp}_{\hat{\mathbf{x}}^s}^{-1}(\mathbf{x}_0^s)\|^3] = \mathbb{E}F(\hat{\mathbf{x}}^s)$, which implies

$$\mathbb{E}F(\hat{\mathbf{x}}^s) - \mathbb{E}F(\hat{\mathbf{x}}^{s+1}) = R_0^s - R_T^s \geq (729\bar{k}^2\sqrt{L_H})^{-1} \sum_{t=1}^T (\mathbb{E}[\mu(\mathbf{x}_t^s)] - \frac{82\delta}{81}).$$

Telescoping the above inequality from $s = 1$ to S yields

$$\Delta_F \geq \sum_{s=1}^S \mathbb{E}F(\hat{\mathbf{x}}^s) - \mathbb{E}F(\hat{\mathbf{x}}^{s+1}) \geq (729\bar{k}^2\sqrt{L_H})^{-1} \sum_{s=1}^S \sum_{t=1}^T (\mathbb{E}[\mu(\mathbf{x}_t^s)] - \frac{82\delta}{81}).$$

By the definition of \mathbf{x}_{out} , the proof is completed. \square

Corollary 4.2. For any s and t , let $\tilde{\mathbf{h}}_t^s$ be an inexact solution of the cubic subproblem $m_t^s(\mathbf{h})$, which satisfies Definition 4.4 with $\delta = (1500\bar{k}^2\sqrt{L_H})^{-1}\epsilon^{3/2}$. Suppose that the cubic regularization parameter σ in Algorithm 1 is fixed and satisfies $\sigma = \bar{k}L_H$, where L_H is the Hessian Lipschitz parameter according to (10) with $\bar{k} \geq 2$, and it also satisfies (61). Let the epoch length $T = N^{1/5}$, batch sizes $b_g = \frac{3000^{4/3}N^{4/5}(\sqrt{\xi-1}+1)^4}{\bar{k}^2}$, $b_h = \frac{e \log d}{(\sqrt{\frac{\bar{k}}{193N^{1/5}(\sqrt{\xi-1}+1)} + \frac{1}{8} - \frac{1}{2\sqrt{2}}})^2}$, and the number of epochs

$S = \max\{1, 1500\bar{k}^2L_H^{1/2}\Delta_F N^{-1/5}\epsilon^{-3/2}\}$. Then, under Assumptions 1-5, Algorithm 1 finds an $(\epsilon, \sqrt{L_H}\epsilon)$ -second-order stationary point in $\tilde{O}(N + L_H^{1/2}\Delta_F N^{4/5}\epsilon^{-3/2})$ number of second-order oracle calls.

Proof. Under the parameter setting in Corollary 4.2 and Theorem 4.2, we have

$$\mathbb{E}[\mu(\mathbf{x}_{out})] \leq \frac{729\bar{k}^2L_H^{1/2}\Delta_F}{ST} + 738\bar{k}^2\sqrt{L_H}\delta \leq \frac{\epsilon^{3/2}}{2} + \frac{\epsilon^{3/2}}{2} = \epsilon^{3/2}. \quad (71)$$

Thus, \mathbf{x}_{out} is an $(\epsilon, \sqrt{L_H \epsilon})$ -approximate local minimum. Similar to the discussion in Corollary 4.1, the total number of SO calls is

$$\begin{aligned} SN + (ST)(b_g + b_h) &\leq N + 1500\bar{k}^2 L_H^{1/2} \Delta_F N^{4/5} \epsilon^{-3/2} + 1500\bar{k}^2 L_H^{1/2} \Delta_F \epsilon^{-3/2} (b_g + b_h) \\ &= \tilde{O}(N + L_H^{1/2} \Delta_F N^{4/5} \epsilon^{-3/2}). \end{aligned}$$

□

5 Numerical Studies

In this section, we conduct numerical experiments to verify our theoretical complexity results for the R-SVRC algorithm to find a second-order stationary point. Besides different simulation studies, we compare our algorithm with crude Riemannian cubic regularization Newton method (CRC), Riemannian adaptive cubic regularization method (ARC), and Riemannian trust region method (RTR) – see Zhang & Zhang (2018), Agarwal *et al.* (2020), Absil *et al.* (2007). Our code is written in conformance with the Manopt package Boumal *et al.* (2014), and it is available at <https://github.com/samdavanloo/R-SVRC>. All the numerical studies are run on a laptop with 1.4 GHz Quad-Core Intel Core i5 CPU and 8 GB memory.

5.1 Parameter Estimation of Multivariate Student’s t-distribution

The maximum likelihood estimation of the (scale) parameter of the multivariate t-distribution (2) requires solving

$$\min_{X \in \mathcal{S}_{++}^p} F(X) = \frac{\nu + p}{2N} \sum_{i=1}^N \log(1 + \frac{\mathbf{a}_i^T X \mathbf{a}_i}{\nu}) - \frac{1}{2} \log \det(X), \quad (72)$$

where the mean is assumed to be zero and X is the inverse of the scale matrix Σ which should belong to the Symmetric Positive Definite (SPD) manifold. The Euclidean gradient and Hessian of F can be calculated as

$$\nabla F(X) = \frac{\nu + p}{2N} \sum_{i=1}^N \frac{\mathbf{a}_i \mathbf{a}_i^T}{\nu + \mathbf{a}_i^T X \mathbf{a}_i} - \frac{1}{2} X^{-1}, \quad (73)$$

$$\nabla^2 F(X)[U] = \frac{\nu + p}{2N} \sum_{i=1}^N \frac{-\mathbf{a}_i^T U \mathbf{a}_i}{(\nu + \mathbf{a}_i^T X \mathbf{a}_i)^2} \mathbf{a}_i \mathbf{a}_i^T + \frac{1}{2} X^{-1} U X^{-1} \quad (74)$$

$$= \left[\frac{\nu + p}{2N} \sum_{i=1}^N \frac{-(\mathbf{a}_i \mathbf{a}_i^T) \otimes (\mathbf{a}_i \mathbf{a}_i^T)}{(\nu + \mathbf{a}_i^T X \mathbf{a}_i)^2} + \frac{1}{2} X^{-1} \otimes X^{-1} \right] \cdot \text{vec}(U), \quad (75)$$

where $\text{sym}(Y) \triangleq \frac{1}{2}(Y + Y^T)$, $\text{vec}(\cdot)$ denotes vectorization of the input matrix, and \otimes denotes the Kronecker product. The Riemannian gradient and Hessian are obtained as (see Bhatia (2009), Boumal *et al.* (2014)):

$$\text{grad} F(X) = X \text{sym}(\nabla F(X)) X, \quad (76)$$

$$\text{Hess} F(X)[U] = X \text{sym}(\nabla^2 F(X)[U]) X + \text{sym}(U \nabla F(X) X). \quad (77)$$

While the above equations compute the full gradient and Hessian along certain direction, the stochastic gradient and Hessian along certain direction also easily follow. For instance, for function

$$F_{I_h}(X) \triangleq \frac{1}{b_h} \sum_{i \in I_h} f_i(X),$$

the second term in the formula for \mathbf{U}_t^s (see Step 9 in Algorithm 1) can be calculated as

$$\nabla^2 F_{I_h}(X)[U] = \left[\frac{\nu + p}{2b_h} \sum_{i \in I_h} \frac{-(\mathbf{a}_i \mathbf{a}_i^T) \otimes (\mathbf{a}_i \mathbf{a}_i^T)}{(\nu + \mathbf{a}_i^T X \mathbf{a}_i)^2} + \frac{1}{2} X^{-1} \otimes X^{-1} \right] \cdot \text{vec}(U), \quad (78)$$

$$\text{Hess} F_{I_h}(X)[U] = X \text{sym}(\nabla^2 F_{I_h}(X)[U])X + \text{sym}(U \nabla F_{I_h}(X)X). \quad (79)$$

While computing the Euclidean gradient and Hessian (along certain direction) using (73) and (75) requires processing N data points, transforming them to their Riemannian counterparts is relatively simple, in the sense that their computation is independent of N . Therefore, at the beginning of each epoch, the tensor inside the square bracket in (75) is computed and stored. In the following within-epoch iterations, to update \mathbf{U}_t^s (Step 9 in Algorithm 1), only the second and third terms need to be updated which can be performed efficiently as the batch size is small compared to N .

5.2 Linear Classifier Over the Sphere Manifold

In this example, we consider a classification problem based on N training examples $\{\mathbf{a}_i, b_i\}_{i=1}^N$ where $\mathbf{a}_i \in \mathbb{R}^m$ and $b_i \in \{-1, 1\}$ for all $i \in [N]$. We aim to estimate the model parameter \mathbf{x} for a linear classifier $f(\mathbf{a}) = \mathbf{x}^\top \mathbf{a}$ such that it minimizes a smooth nonconvex loss function Zhao *et al.* (2010), Li & Yang (2003)

$$\mathcal{L}(\mathbf{x}; \{(\mathbf{a}_i, b_i)\}_{i=1}^N) = \sum_{i=1}^N \left(1 - \frac{1}{1 + e^{-b_i \cdot \mathbf{x}^\top \mathbf{a}_i}}\right)^2, \quad (80)$$

over the Sphere manifold, $\{\mathbf{x} \in \mathbb{R}^m : \mathbf{x}^\top \mathbf{x} = 1\}$ Absil *et al.* (2009). The Euclidean gradient and Hessian of \mathcal{L} are

$$\begin{aligned} \nabla \mathcal{L}(\mathbf{x}) &= \sum_{i=1}^N -\frac{e^{-2b_i(\mathbf{x}^\top \mathbf{a}_i)}}{(1 + e^{-b_i(\mathbf{x}^\top \mathbf{a}_i)})^3} \mathbf{a}_i, \\ \nabla^2 \mathcal{L}(\mathbf{x}) &= \sum_{i=1}^N \frac{(2 - e^{-b_i \mathbf{x}^\top \mathbf{a}_i}) b_i^2 e^{-2b_i(\mathbf{x}^\top \mathbf{a}_i)}}{(1 + e^{-b_i(\mathbf{x}^\top \mathbf{a}_i)})^4} \mathbf{a}_i \mathbf{a}_i^\top. \end{aligned}$$

The Riemannian gradient and Hessian of \mathcal{L} along U (see Proposition 5.3.2 in Absil *et al.* (2009), Boumal *et al.* (2014)) are

$$\text{grad} \mathcal{L}(\mathbf{x}) = P_{\mathbf{x}}(\nabla \mathcal{L}(\mathbf{x})), \quad (81)$$

$$\text{Hess} \mathcal{L}(\mathbf{x})[\mathbf{u}] = P_{\mathbf{x}}(\nabla^2 \mathcal{L}(\mathbf{x})[\mathbf{u}]) - (\mathbf{x}^\top \nabla \mathcal{L}(\mathbf{x})) \mathbf{u}, \quad (82)$$

where the tangent space projection is $P_{\mathbf{x}}(\mathbf{y}) \triangleq \mathbf{y} - (\mathbf{x}^\top \mathbf{y}) \mathbf{x}$. The stochastic gradient and Hessian easily follows by summing the corresponding terms over the minibatch.

The first example above on estimating the inverse scale matrix of the multivariate t-distribution over symmetric positive definite (SPD) satisfies Assumptions 3-5, and its objective function satisfies

Assumptions 1-2 if the minimum eigenvalue of the matrices is bounded away from zero. The second example on estimating the parameter of the linear classifier over sphere manifold satisfies all of the assumptions, i.e., the sphere manifold satisfies Assumptions 3-5 and Assumptions 1-2 follows from continuous differentiability of the objective function and compactness of the sphere manifold.

5.3 Numerical Results

Data Simulation. The first numerical study is to estimate the inverse scale (covariance) matrix of the multivariate Student’s t-distribution (see problem (72)). Data is simulated from a multivariate t-distribution with three degrees of freedom and randomly generate scale matrix $\Sigma_{\text{true}} \in \mathcal{S}_{++}^d$ with $d = 10$. $N = 10^4$ samples are generated from the underlying distribution which are then added with the Gaussian noise ϵ sampled from $\mathcal{N}(0, \tau^2 \mathbf{I}_d)$ with τ^2 equal to 0.1, 1, 5, and 10.

The second numerical study is to estimate the parameter of a linear classifier over the Sphere manifold (see problem (80)). To simulate the data, the true parameter \mathbf{x}_{true} is first generated from $\mathcal{N}(0, \mathbf{I}_d)$ which is then normalized to belong to the Sphere manifold. Next, $\mathbf{a}_i \in \mathbb{R}^{d \times 1}$, $i = 1, \dots, N$ are randomly generated from the uniform distribution where $d = 20$ and $N = 10^5$. The corresponding label b_i to \mathbf{a}_i is set to 1 if $\mathbf{x}_{\text{true}}^\top \mathbf{a}_i + \epsilon_i > 0$, where $\epsilon_i \sim \mathcal{N}(0, \tau^2)$, and -1 , otherwise, where τ^2 is chosen to be 0.02, 0.1, 1, and 3.

The proposed R-SVRC algorithm is run 15 times in each numerical study. The shaded plots discussed in the Results below provide percentile information based on these replicates.

Number of calls to the stochastic oracle. For the R-SVRC method, at the beginning of each epoch, the SO is called N times. However, within each epoch, each iteration makes $(b_g + b_h)$ calls to SO. In the deterministic CRC, ARC and RTR methods, each iteration makes N calls to SO. The number of SO calls and the CPU runtime are the two performance measures we have used to compare the proposed method with the other second-order methods.

Parameters and subproblem solver. The g-smoothness and H-smoothness assumptions is standard in nonasymptotic analysis in Riemannian optimization - see, e.g., Absil *et al.* (2004, 2009), Boumal *et al.* (2019), Boumal (2020). However, obtaining the g-smoothness and H-smoothness constants is not trivial and we defer it to future studies. In the following, we numerically analyze the effect of different parameters on the performance of Algorithm 1, i.e., epoch size T , cubic regularization constant σ , batchsize b_g and b_h . The cubic subproblem (Step 9 of the Algorithm 1) is solved using the conjugated gradient method using the Manopt solver Boumal *et al.* (2014).

To estimate the inverse scale matrix of the multivariate t-distribution over the symmetric positive definite manifold, the default optimization parameter setting for Algorithm 1 is $\sigma = 0.01$, $b_g = b_h = 500$ and $T = 5$. To estimate the parameter of the linear classifier over Sphere manifold, the default optimization parameter setting for Algorithm 1 is $\sigma = 0.1$, $b_g = b_h = 5000$ and $T = 5$.

Results. Figure 1 shows the performance of the R-SVRC algorithm for different levels of noise ϵ added to the simulated data (see data simulation above). The top two plots in Figure 1 show the proposed algorithm successfully approach a second-order stationary point in all scenarios. As the output of Algorithm 1 is to be sampled uniformly at random for $s \in [S]$ and $t \in [T]$, we also plot the averaged $\mu(\mathbf{x}^k)$ sequence (over iterations) in the bottom two plots. These averaged sequences show $\mathbb{E}(\mu(\mathbf{x}^k))$ decreases with a sublinear rate which is consistent with the first main theorem.

Figure 1: Effect of the added noise to the simulated data on the performance of the proposed R-SVRC algorithm over 15 replicates. **(Left)** Estimating inverse scale matrix of multivariate t-distribution over SPD manifold. **(Right)** Estimating parameter of the linear classifier over Sphere manifold.

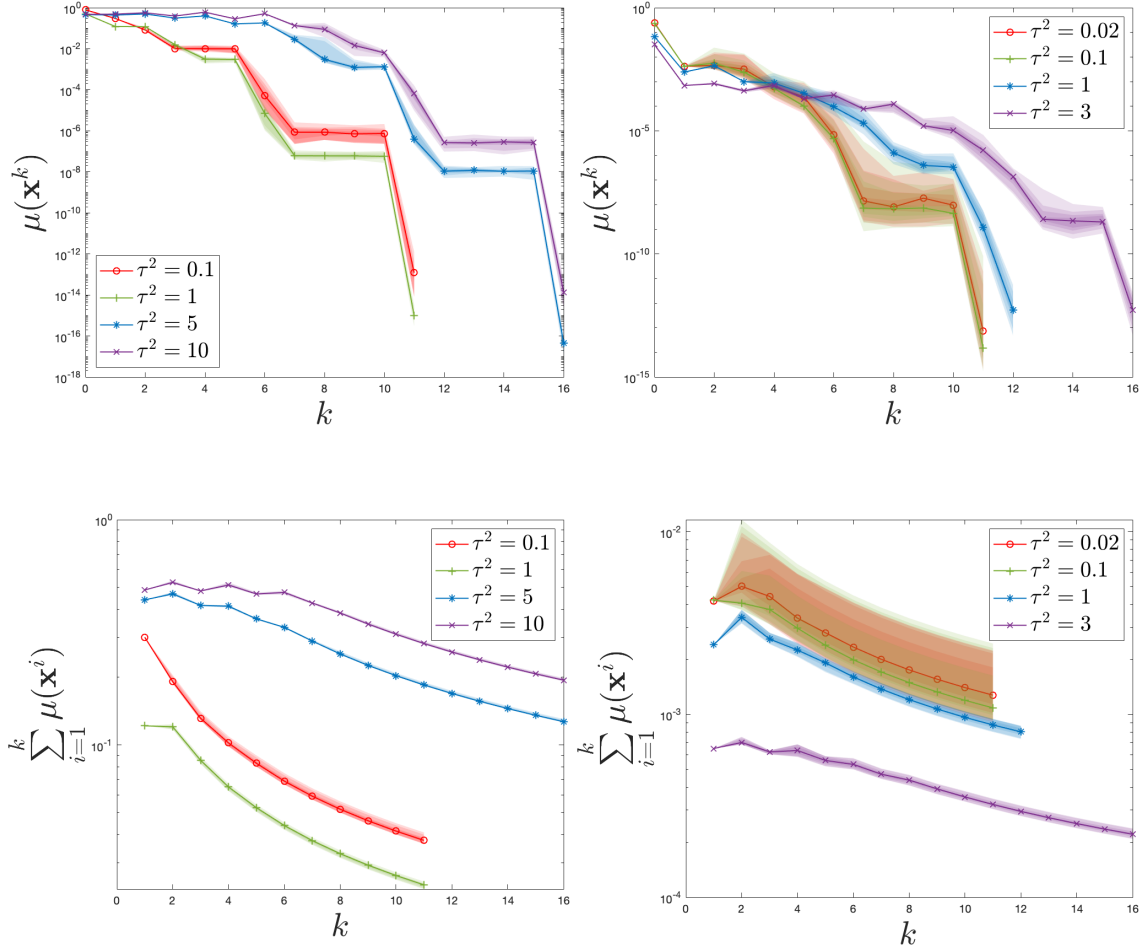


Figure 2: Performance of the proposed R-SVRC algorithm for different optimization parameter settings over 15 replicates. **(Left)** Estimating inverse scale matrix of multivariate t-distribution over SPD manifold. **(Right)** Estimating parameter of the linear classifier over Sphere manifold.

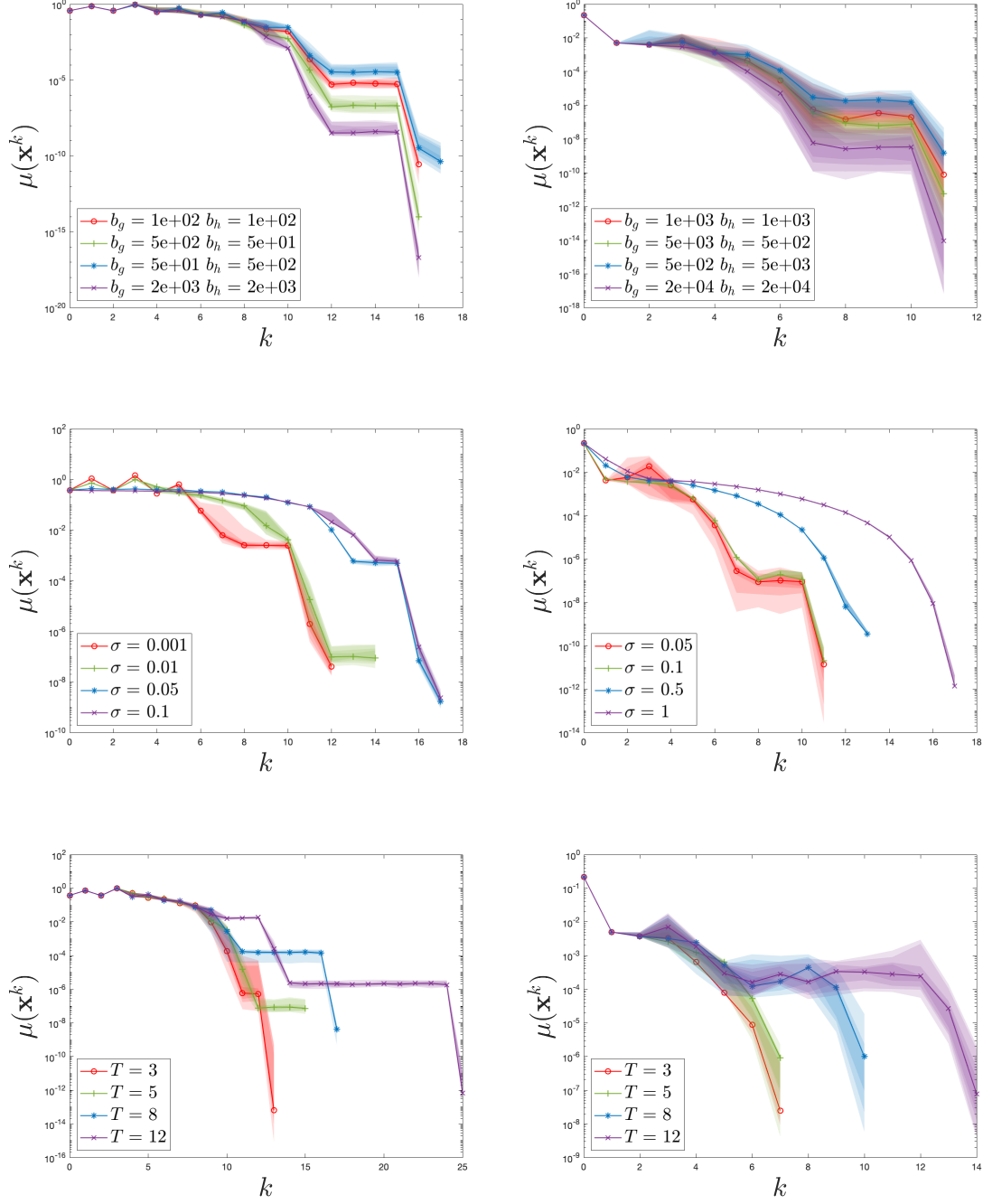


Figure 2 illustrates the performance of the R-SVRC algorithm for both numerical studies for different settings of the optimization parameters. Left and right columns corresponds to the first and second numerical studies, respectively. Most of the plots show a superlinear rate of convergence to a second-order stationary point using the last iterate as the output of the algorithm. The first row shows that smaller batch sizes result in slower convergence with early oscillation around the plateau. Specifically, the top three lines have ascending values of b_g while descending values of b_h which implies that the effect of b_g is more significant than that of b_h . The second row shows that bigger values of σ can provide smaller objective values but with slower convergence rate. Furthermore, larger σ values tends to produce a smaller $\|\mathbf{h}\|$ based on the subproblem (27) which leads to more stable and smooth sequences shown in the plots. The third row shows that bigger values of T , i.e., less frequent full gradient and Hessian calculations, result in slower rate of convergence for a fixed number of iterations which is also intuitive.

In Figure 3, we compare the proposed R-SVRC method with the other three benchmark methods, Riemannian adaptive cubic regularization method (ARC), Riemannian trust region method (RTR) and crude Riemannian cubic regularization method (CRC) Agarwal *et al.* (2020), Boumal (2015), Zhang & Zhang (2018) over the number stochastic oracle calls (see Definition 4.3) and also cpu time. Results show faster decrease by the R-SVRC method compared to the other benchmark methods.

Finally, Figure 4 visualizes the optimization path obtained by the R-SVRC algorithm over the Sphere manifold. The generated iterates converge to the optimal solution.

6 Conclusions

We developed the Riemannian stochastic variance-reduced cubic-regularized Newton method (R-SVRC) for optimization over Riemannian manifolds embedded in a Euclidean space. The proposed double-loop algorithm requires information on the full gradient and Hessian at the beginning of each epoch (outer loop) but updates the gradient and Hessian within each epoch in a stochastic variance-reduced fashion. Each iteration requires solving a cubic-regularized Newton subproblem. Iteration complexity of the proposed algorithm to find a second-order stationary points is established which matches the worst-case complexity bounds in the Euclidean setting. Furthermore, a version of the algorithm which only requires an inexact solution to the cubic regularized Newton subproblem is proposed which has the same complexity bound as the exact case. Finally, the performance of the proposed algorithm is evaluated over two numerical studies with symmetric positive definite and sphere manifolds.

References

- Absil, P-A, & Hosseini, Seyedehsomyayeh. 2019. A collection of nonsmooth Riemannian optimization problems. *Pages 1–15 of: Nonsmooth Optimization and Its Applications*. Springer.
- Absil, P-A, Baker, Christopher G, & Gallivan, Kyle A. 2007. Trust-region methods on Riemannian manifolds. *Foundations of Computational Mathematics*, **7**(3), 303–330.
- Absil, P-A, Mahony, Robert, & Sepulchre, Rodolphe. 2009. *Optimization algorithms on matrix manifolds*. Princeton University Press.
- Absil, Pierre-Antoine, Baker, Christopher G, & Gallivan, Kyle A. 2004. Trust-region methods on Riemannian manifolds with applications in numerical linear algebra. *Pages 5–9 of: Proceed-*

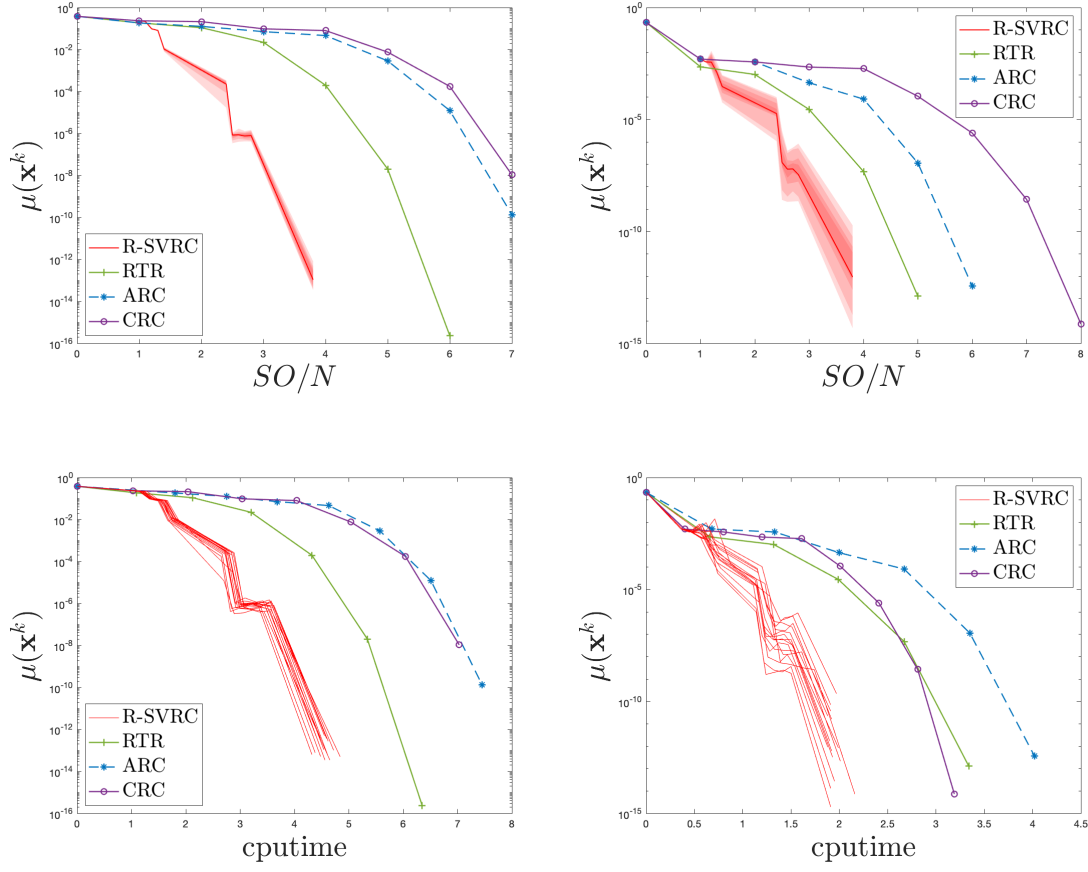


Figure 3: Performance of the proposed R-SVRC algorithm compared to the three benchmark methods. **(Left)** Estimating inverse scale matrix of multivariate t-distribution over SPD manifold. **(Right)** Estimating parameter of the linear classifier over Sphere manifold.

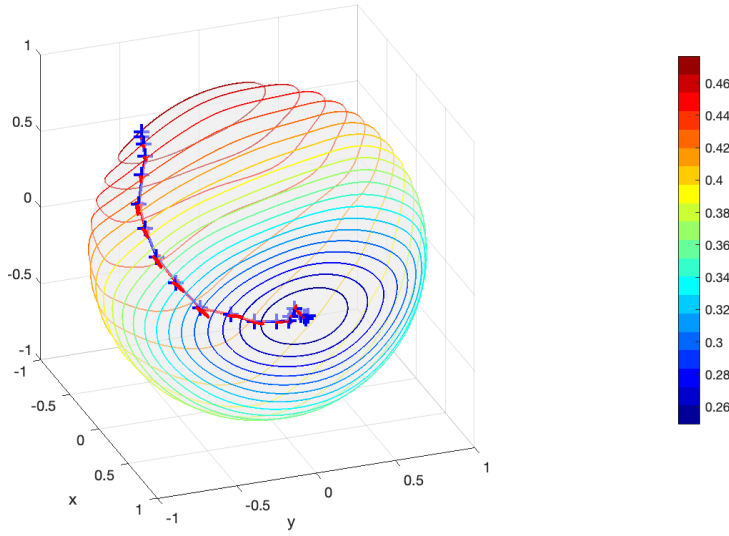


Figure 4: The path of the R-SVRC algorithm iterates over the 2-dimensional instance of the Sphere manifold in the second numerical study.

ings of the 16th International Symposium on Mathematical Theory of Networks and Systems (MTNS2004), Leuven, Belgium.

- Agarwal, Naman, Boumal, Nicolas, Bullins, Brian, & Cartis, Coralia. 2018. *Adaptive regularization with cubics on manifolds*.
- Agarwal, Naman, Boumal, Nicolas, Bullins, Brian, & Cartis, Coralia. 2020. Adaptive regularization with cubics on manifolds. *Mathematical Programming*, 1–50.
- Ahn, Kwangjun, & Sra, Suvrit. 2020. From Nesterov’s estimate sequence to Riemannian acceleration. *Pages 84–118 of: Conference on Learning Theory*. PMLR.
- Alimisis, Foivos, Orvieto, Antonio, Bécigneul, Gary, & Lucchi, Aurelien. 2020. A continuous-time perspective for modeling acceleration in Riemannian optimization. *Pages 1297–1307 of: International Conference on Artificial Intelligence and Statistics*. PMLR.
- Alimisis, Foivos, Orvieto, Antonio, Bécigneul, Gary, & Lucchi, Aurelien. 2021. Momentum Improves Optimization on Riemannian Manifolds. *Pages 1351–1359 of: International Conference on Artificial Intelligence and Statistics*. PMLR.
- Arjovsky, Martin, Shah, Amar, & Bengio, Yoshua. 2016. Unitary evolution recurrent neural networks. *Pages 1120–1128 of: International Conference on Machine Learning*.
- Baker, Christopher G, Absil, P-A, & Gallivan, Kyle A. 2008. An implicit trust-region method on Riemannian manifolds. *IMA journal of numerical analysis*, **28**(4), 665–689.
- Bansal, Nitin, Chen, Xiaohan, & Wang, Zhangyang. 2018. Can we gain more from orthogonality regularizations in training deep CNNs? *arXiv preprint arXiv:1810.09102*.

- Bento, GC, Ferreira, OP, & Oliveira, PR. 2015. Proximal point method for a special class of nonconvex functions on Hadamard manifolds. *Optimization*, **64**(2), 289–319.
- Bento, Glaydston C, Ferreira, Orizon P, & Melo, Jefferson G. 2017. Iteration-complexity of gradient, subgradient and proximal point methods on Riemannian manifolds. *Journal of Optimization Theory and Applications*, **173**(2), 548–562.
- Bhatia, Rajendra. 2009. *Positive definite matrices*. Princeton university press.
- Bonnabel, Silvere. 2013. Stochastic gradient descent on Riemannian manifolds. *IEEE Transactions on Automatic Control*, **58**(9), 2217–2229.
- Boumal, N., Mishra, B., Absil, P.-A., & Sepulchre, R. 2014. Manopt, a Matlab Toolbox for Optimization on Manifolds. *Journal of Machine Learning Research*, **15**(42), 1455–1459.
- Boumal, Nicolas. 2015. Riemannian trust regions with finite-difference Hessian approximations are globally convergent. *Pages 467–475 of: International Conference on Geometric Science of Information*. Springer.
- Boumal, Nicolas. 2020. An introduction to optimization on smooth manifolds. *Available online, Aug.*
- Boumal, Nicolas, Absil, Pierre-Antoine, & Cartis, Coralia. 2019. Global rates of convergence for nonconvex optimization on manifolds. *IMA Journal of Numerical Analysis*, **39**(1), 1–33.
- Carmon, Yair, & Duchi, John. 2019. Gradient descent finds the cubic-regularized nonconvex Newton step. *SIAM Journal on Optimization*, **29**(3), 2146–2178.
- Cartis, Coralia, Gould, Nicholas IM, & Toint, Philippe L. 2011a. Adaptive cubic regularisation methods for unconstrained optimization. Part I: motivation, convergence and numerical results. *Mathematical Programming*, **127**(2), 245–295.
- Cartis, Coralia, Gould, Nicholas IM, & Toint, Philippe L. 2011b. Adaptive cubic regularisation methods for unconstrained optimization. Part II: worst-case function-and derivative-evaluation complexity. *Mathematical programming*, **130**(2), 295–319.
- Cartis, Coralia, Gould, Nicholas IM, & Toint, Ph L. 2012. Complexity bounds for second-order optimality in unconstrained optimization. *Journal of Complexity*, **28**(1), 93–108.
- Cartis, Coralia, Gould, Nicholas IM, & Toint, Philippe L. 2014. On the complexity of finding first-order critical points in constrained nonlinear optimization. *Mathematical Programming*, **144**(1), 93–106.
- Chavel, Isaac. 2006. *Riemannian geometry: a modern introduction*. Vol. 98. Cambridge university press.
- Cogswell, Michael, Ahmed, Faruk, Girshick, Ross, Zitnick, Larry, & Batra, Dhruv. 2015. Reducing overfitting in deep networks by decorrelating representations. *arXiv preprint arXiv:1511.06068*.
- Criscitiello, Chris, & Boumal, Nicolas. 2020. An accelerated first-order method for non-convex optimization on manifolds. *arXiv preprint arXiv:2008.02252*.
- Criscitiello, Christopher, & Boumal, Nicolas. 2019. Efficiently escaping saddle points on manifolds. *Pages 5987–5997 of: Advances in Neural Information Processing Systems*.
- da Cruz Neto, JX, De Lima, LL, & Oliveira, PR. 1998. Geodesic algorithms in Riemannian geometry. *Balkan J. Geom. Appl*, **3**(2), 89–100.

- de Carvalho Bento, Glaydston, da Cruz Neto, João Xavier, & Oliveira, Paulo Roberto. 2016. A new approach to the proximal point method: convergence on general Riemannian manifolds. *Journal of Optimization Theory and Applications*, **168**(3), 743–755.
- de Melo Mendes, Beatriz Vaz, & de Souza, Rafael Martins. 2004. Measuring financial risks with copulas. *International Review of Financial Analysis*, **13**(1), 27–45.
- Domino, Krzysztof. 2018. Selected Methods for non-Gaussian Data Analysis. *arXiv preprint arXiv:1811.10486*.
- Durrett, Rick. 2019. *Probability: theory and examples*. Vol. 49. Cambridge university press.
- Ferreira, OP, & Oliveira, PR. 2002. Proximal point algorithm on Riemannian manifolds. *Optimization*, **51**(2), 257–270.
- Ferreira, Orizon P, Louzeiro, Mauricio S, & Prudente, LF4018420. 2019. Gradient method for optimization on Riemannian manifolds with lower bounded curvature. *SIAM Journal on Optimization*, **29**(4), 2517–2541.
- Gabay, Daniel. 1982. Minimizing a differentiable function over a differential manifold. *Journal of Optimization Theory and Applications*, **37**(2), 177–219.
- Hosseini, Reshad, & Sra, Suvrit. 2020. Recent advances in stochastic Riemannian optimization. *Handbook of Variational Methods for Nonlinear Geometric Data*, 527–554.
- Hu, Jiang, Milzarek, Andre, Wen, Zaiwen, & Yuan, Yaxiang. 2018. Adaptive quadratically regularized Newton method for Riemannian optimization. *SIAM Journal on Matrix Analysis and Applications*, **39**(3), 1181–1207.
- Hu, Jiang, Liu, Xin, Wen, Zai-Wen, & Yuan, Ya-Xiang. 2020. A brief introduction to manifold optimization. *Journal of the Operations Research Society of China*, **8**(2), 199–248.
- Huang, Lei, Liu, Xianglong, Lang, Bo, Yu, Adams Wei, Wang, Yongliang, & Li, Bo. 2018. Orthogonal weight normalization: Solution to optimization over multiple dependent Stiefel manifolds in deep neural networks. In: *Thirty-Second AAAI Conference on Artificial Intelligence*.
- Huang, Wen, & Wei, Ke. 2021. Riemannian proximal gradient methods. *Mathematical Programming*, 1–43.
- Jin, Chi, Netrapalli, Praneeth, Ge, Rong, Kakade, Sham M, & Jordan, Michael I. 2019. Stochastic gradient descent escapes saddle points efficiently. *arXiv preprint arXiv:1902.04811*.
- Johnson, Rie, & Zhang, Tong. 2013. Accelerating stochastic gradient descent using predictive variance reduction. *Advances in neural information processing systems*, **26**, 315–323.
- Kasai, Hiroyuki, & Mishra, Bamdev. 2018. Inexact trust-region algorithms on Riemannian manifolds. *Pages 4254–4265 of: NeurIPS*.
- Kasai, Hiroyuki, Sato, Hiroyuki, & Mishra, Bamdev. 2017. Riemannian stochastic quasi-Newton algorithm with variance reduction and its convergence analysis. *arXiv preprint arXiv:1703.04890*.
- Kasai, Hiroyuki, Sato, Hiroyuki, & Mishra, Bamdev. 2018. Riemannian stochastic recursive gradient algorithm. *Pages 2516–2524 of: International Conference on Machine Learning*. PMLR.
- Kotz, Samuel, & Nadarajah, Saralees. 2004. *Multivariate t-distributions and their applications*. Cambridge University Press.
- Kovalev, Dmitry, Mishchenko, Konstantin, & Richtárik, Peter. 2019. Stochastic Newton and Cubic Newton Methods with Simple Local Linear-Quadratic Rates. *arXiv preprint arXiv:1912.01597*.

- Krzanowski, Wojtek J, & FHC, Marriott. 1994. *Multivariate analysis*. Wiley.
- Lee, John M. 2018. *Introduction to Riemannian manifolds*. Springer.
- Li, Fan, & Yang, Yiming. 2003. A loss function analysis for classification methods in text categorization. *Pages 472–479 of: Proceedings of the 20th international conference on machine learning (ICML-03)*.
- Li, Jun, Fuxin, Li, & Todorovic, Sinisa. 2020. Efficient Riemannian optimization on the Stiefel manifold via the Cayley transform. *arXiv preprint arXiv:2002.01113*.
- Liu, Yuanyuan, Shang, Fanhua, Cheng, James, Cheng, Hong, & Jiao, Licheng. 2017. Accelerated First-order Methods for Geodesically Convex Optimization on Riemannian Manifolds. *Pages 4868–4877 of: NIPS*.
- Luenberger, David G. 1972. The gradient projection method along geodesics. *Management Science*, **18**(11), 620–631.
- Mackey, Lester, Jordan, Michael I, Chen, Richard Y, Farrell, Brendan, Tropp, Joel A, *et al*. 2014. Matrix concentration inequalities via the method of exchangeable pairs. *The Annals of Probability*, **42**(3), 906–945.
- Nesterov, Yurii, & Polyak, Boris T. 2006. Cubic regularization of Newton method and its global performance. *Mathematical Programming*, **108**(1), 177–205.
- Nguyen, Lam M, Liu, Jie, Scheinberg, Katya, & Takáč, Martin. 2017. Stochastic recursive gradient algorithm for nonconvex optimization. *arXiv preprint arXiv:1705.07261*.
- Nocedal, Jorge, & Wright, Stephen. 2006. *Numerical optimization*. Springer Science & Business Media.
- Qi, Chunhong. 2011. *Numerical optimization methods on Riemannian manifolds*. Ph.D. thesis, Florida State University.
- Ring, Wolfgang, & Wirth, Benedikt. 2012. Optimization methods on Riemannian manifolds and their application to shape space. *SIAM Journal on Optimization*, **22**(2), 596–627.
- Roychowdhury, Anirban. 2017. Accelerated stochastic quasi-Newton optimization on Riemann manifolds. *arXiv preprint arXiv:1704.01700*.
- Rudin, Walter, *et al*. 1964. *Principles of mathematical analysis*. Vol. 3. McGraw-hill New York.
- Ruszczynski, Andrzej. 2011. *Nonlinear optimization*. Princeton university press.
- Sato, Hiroyuki. 2021. *Riemannian Optimization and Its Applications*. Springer Nature.
- Sato, Hiroyuki, & Iwai, Toshihiro. 2015. A new, globally convergent Riemannian conjugate gradient method. *Optimization*, **64**(4), 1011–1031.
- Sato, Hiroyuki, Kasai, Hiroyuki, & Mishra, Bamdev. 2019. Riemannian stochastic variance reduced gradient algorithm with retraction and vector transport. *SIAM Journal on Optimization*, **29**(2), 1444–1472.
- Smith, Steven T. 1994. Optimization techniques on Riemannian manifolds. *Fields institute communications*, **3**(3), 113–135.
- Smith, Steven Thomas. 1993. *Geometric optimization methods for adaptive filtering*. Harvard University.
- Sra, Suvrit, & Hosseini, Reshad. 2015. Conic geometric optimization on the manifold of positive definite matrices. *SIAM Journal on Optimization*, **25**(1), 713–739.

- Sun, Yifan, Zheng, Liang, Deng, Weijian, & Wang, Shengjin. 2017. Svdnet for pedestrian retrieval. *Pages 3800–3808 of: Proceedings of the IEEE international conference on computer vision*.
- Sun, Yue, Flammarion, Nicolas, & Fazel, Maryam. 2019. Escaping from saddle points on Riemannian manifolds. *Pages 7276–7286 of: Advances in Neural Information Processing Systems*.
- Szegö, Giorgio. 2002. Measures of risk. *Journal of Banking & finance*, **26**(7), 1253–1272.
- Tripuraneni, Nilesh, Flammarion, Nicolas, Bach, Francis, & Jordan, Michael I. 2018. Averaging stochastic gradient descent on Riemannian manifolds. *Pages 650–687 of: Conference On Learning Theory*. PMLR.
- Udriste, Constantin. 2013. *Convex functions and optimization methods on Riemannian manifolds*. Vol. 297. Springer Science & Business Media.
- Wisdom, Scott, Powers, Thomas, Hershey, John, Le Roux, Jonathan, & Atlas, Les. 2016. Full-capacity unitary recurrent neural networks. *Pages 4880–4888 of: Advances in neural information processing systems*.
- Xie, Di, Xiong, Jiang, & Pu, Shiliang. 2017. All you need is beyond a good init: Exploring better solution for training extremely deep convolutional neural networks with orthonormality and modulation. *Pages 6176–6185 of: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
- Zhang, Hongyi, & Sra, Suvrit. 2016. First-order methods for geodesically convex optimization. *Pages 1617–1638 of: Conference on Learning Theory*. PMLR.
- Zhang, Hongyi, & Sra, Suvrit. 2018. Towards Riemannian accelerated gradient methods. *arXiv preprint arXiv:1806.02812*.
- Zhang, Hongyi, Reddi, Sashank J, & Sra, Suvrit. 2016. Riemannian SVRG: Fast stochastic optimization on Riemannian manifolds. *Pages 4592–4600 of: Advances in Neural Information Processing Systems*.
- Zhang, Junyu, & Zhang, Shuzhong. 2018. A Cubic Regularized Newton’s Method over Riemannian Manifolds. *arXiv preprint arXiv:1805.05565*.
- Zhao, Lei, Mammadov, Musa, & Yearwood, John. 2010. From convex to nonconvex: a loss function analysis for binary classification. *Pages 1281–1288 of: 2010 IEEE International Conference on Data Mining Workshops*. IEEE.
- Zhou, Dongruo, Xu, Pan, & Gu, Quanquan. 2018. Stochastic variance-reduced cubic regularized Newton methods. *Pages 5990–5999 of: International Conference on Machine Learning*. PMLR.