SIAM J. OPTIM.
Vol. 29, No. 2, pp. 1444–1472

RIEMANNIAN STOCHASTIC VARIANCE REDUCED GRADIENT ALGORITHM WITH RETRACTION AND VECTOR TRANSPORT*

HIROYUKI SATO†, HIROYUKI KASAI‡, AND BAMDEV MISHRA§

Abstract. In recent years, stochastic variance reduction algorithms have attracted considerable attention for minimizing the average of a large but finite number of loss functions. This paper proposes a novel Riemannian extension of the Euclidean stochastic variance reduced gradient (R-SVRG) algorithm to a manifold search space. The key challenges of averaging, adding, and subtracting multiple gradients are addressed with retraction and vector transport. For the proposed algorithm, we present a global convergence analysis with a decaying step size as well as a local convergence rate analysis with a fixed step size under some natural assumptions. In addition, the proposed algorithm is applied to the computation problem of the Riemannian centroid on the symmetric positive definite (SPD) manifold as well as the principal component analysis and low-rank matrix completion problems on the Grassmann manifold. The results show that the proposed algorithm outperforms the standard Riemannian stochastic gradient descent algorithm in each case.

Key words. Riemannian optimization, stochastic variance reduced gradient, retraction, vector transport, Riemannian centroid, principal component analysis, matrix completion

AMS subject classifications. 90C06, 90C15, 90C30

DOI. 10.1137/17M1116787

1. Introduction.

A general loss minimization problem is defined as $\min_w f(w)$, where $f(w) := \frac{1}{N}\sum_{n=1}^{N} f_n(w)$, w is the model variable, N is the number of samples, and $f_n(w)$ is the loss incurred on the nth sample. The *full gradient descent* (GD) algorithm requires the evaluation of N derivatives, i.e., $\sum_{n=1}^{N} \nabla f_n(w)$, per iteration, which is computationally expensive when N is extremely large. A well-known alternative uses only one derivative $\nabla f_n(w)$ per iteration for the nth sample, and it forms the basis of the *stochastic gradient descent* (SGD) algorithm. When a relatively large step size is used in SGD, the training loss first decreases rapidly but results in large fluctuations around the solution. Conversely, when a small step size is used, a large number of iterations are required for SGD to converge. To circumvent this problem, SGD starts with a relatively large step size and gradually decreases it.

Recently, *variance reduction* techniques have been proposed to accelerate SGD convergence [7, 15, 21, 27, 29, 30, 32]. The stochastic variance reduced gradient (SVRG) algorithm is a popular technique with excellent convergence properties [15]. For smooth and strongly convex functions, SVRG has convergence rates similar to those of the stochastic dual coordinate ascent algorithm [30] and the stochastic average gradient (SAG) algorithm [27]. Garber and Hazan [9] analyzed the convergence rate of SVRG when f is a convex function that is the sum of nonconvex (but smooth) terms, and they applied their result to the principal component analysis

*Received by the editors February 15, 2017; accepted for publication (in revised form) February 7, 2019; published electronically May 28, 2019.
http://www.siam.org/journals/siopt/29-2/M111678.html
Funding: This study was supported in part by grants JP16K17647 and JP16K00031 from the Grants-in-Aid for Scientific Research Program (KAKENHI) of the Japan Society for the Promotion of Science (JSPS) and by the Kyoto University Hakubi Project.
†Department of Applied Mathematics and Physics, Kyoto University, Yoshida-honmachi, Sakyo-ku, Kyoto 606-8501, Japan (hsato@amp.i.kyoto-u.ac.jp).
‡Graduate School of Informatics and Engineering, The University of Electro-Communications, 1-5-1 Chofugaoka, Chofu, Tokyo 182-8585, Japan (kasai@is.uec.ac.jp).
§Microsoft, Hyderabad, Telangana 500032, India (bamdevm@microsoft.com).

(PCA) problem. Shalev-Shwartz [28] also obtained similar results. Allen-Zhu and Yuan [3] further studied the same case with better convergence rates. Shamir [31] specifically studied the convergence properties of the variance reduction PCA algorithm. More recently, Allen-Zhu and Hazan [2] and Reddi et al. [26] independently proposed variance reduction methods for faster nonconvex optimization. However, it should be noted that all these cases assume a Euclidean search space.

In this paper, we handle problems in which the variables have a manifold structure:

$$\min_{w \in \mathcal{M}} f(w) := \frac{1}{N} \sum_{n=1}^{N} f_n(w), \tag{1}$$

where $\mathcal{M}$ is a Riemannian manifold and f_n, $n = 1, 2, \ldots, N$, are real-valued functions on $\mathcal{M}$. These problems include, e.g., the low-rank matrix completion problem [23], the Riemannian centroid computation problem, and the PCA problem. In all these problems, optimization on *Riemannian manifolds* has shown state-of-the-art performance. The Riemannian framework exploits the geometry of the search space, which is characterized by the constraints of the optimization problem. Numerous efficient optimization algorithms have been developed [1]. Specifically, the problem $\min_{w \in \mathcal{M}} f(w)$, where $\mathcal{M}$ is a Riemannian manifold, is solved as an *unconstrained optimization problem* defined over the Riemannian manifold search space. Furthermore, although these algorithms mainly address *batch-based* approaches, Bonnabel [5] proposed a *Riemannian stochastic gradient descent* (R-SGD) algorithm that extends SGD from Euclidean space to Riemannian manifolds. Recently, more advanced stochastic optimization algorithms have also been generalized to Riemannian manifolds, including R-SQN-VR [18] and R-SRG [17].

Building upon the work of Bonnabel [5], we propose an extension of the SVRG algorithm to a Riemannian manifold search space (R-SVRG) and novel analyses. This extension is nontrivial and requires particular consideration for handling the averaging, addition, and subtraction of multiple gradients at different points on the manifold $\mathcal{M}$. Toward this end, this paper specifically leverages the notions of retraction and vector transport. The algorithm and convergence analysis presented in this paper are generalized, which is in itself a challenging problem, in the retraction and vector transport case, as well as in the exponential mapping and parallel translation case, allowing extremely efficient implementation and making distinct contributions compared with an existing approach [34] that relies only on the exponential mapping and parallel translation case.

It should be mentioned that the recent study [34] by Zhang et al., which appeared simultaneously with our technical report [16], has also proposed R-SVRG on manifolds. The main difference between our work and [34] is that we provide convergence analyses for the algorithm with retraction and vector transport, whereas [34] deals with a special case in which exponential mapping and parallel translation are used as retraction and vector transport, respectively. There are additional differences. Our convergence analysis handles global and local convergence analyses separately, as in the typical analyses of batch algorithms on Riemannian manifolds [1]. Our assumptions for the local convergence rate analysis are imposed only in a local neighborhood around a minimum; they are milder and more natural than the assumptions in [34], which assumes Lipschitz smoothness in the entire space. In other words, our global convergence analysis is not for a convergence rate or complexity, but it is an asymptotic convergence analysis. Here, according to classical usage in nonlinear programming,

we use the term global convergence for convergence to a critical point from any initial point. On the other hand, our local convergence analysis is for a strongly convex function near the optimum, and the class of such functions includes many nonconvex functions that do not necessarily have global strong convexity. Consequently, our analysis should be applicable to different types of manifolds. For example, the parallel translation on the Stiefel manifold, which is an extremely important manifold in practice, is not available in a closed form. We can use a vector transport based on the orthogonal projection to the tangent space of the Stiefel manifold as an efficient implementation. Therefore, compared to the results of [34], our convergence results with retraction and vector transport enable us to deal with a wider variety of manifolds.

We emphasize that, although we derive a local convergence rate with retraction and vector transport in this paper, it can immediately be used to derive a global iteration complexity if we assume, e.g., strong convexity of f globally on the search space. Here, local convergence rate and global iteration complexity imply the rate at which the iterates approach a critical point in a sufficiently small neighborhood and the iteration complexity of obtaining a critical point from an arbitrary initial point, respectively. The result thus obtained can be regarded as a generalization of the global iteration complexity with strongly convex f obtained in [34], where exponential mapping and parallel translation are used.

The remainder of this paper is organized as follows. Section 2 discusses Riemannian optimization theory, including the background on Riemannian geometry and some geometric tools used for optimization on Riemannian manifolds. Section 3 provides a detailed description of R-SVRG. Sections 4 and 5 present the global convergence analysis and local convergence rate analysis of the proposed R-SVRG algorithm, respectively. Section 6 highlights the superior performance of R-SVRG through numerical comparisons with R-SGD on three problems.

The proposed R-SVRG algorithm is implemented in the MATLAB Manopt Toolbox [6]. The MATLAB code for all the proposed algorithms is available at https://github.com/hiroyuki-kasai/RSOpt.

2. Riemannian optimization. Optimization on Riemannian manifolds, also called *Riemannian optimization*, seeks a critical point of a given real-valued function, called the objective or cost function, defined on a smooth Riemannian manifold $\mathcal{M}$. One of the advantages of using Riemannian geometry tools is that the intrinsic properties of the manifold enable us to handle constrained optimization problems as unconstrained optimization problems. This section introduces optimization on manifolds by summarizing [1]. Readers may refer to the various references therein as well as those in [22, 24] for further details.

Let $f\colon \mathcal{M} \to \mathbb{R}$ be a smooth real-valued function on manifold $\mathcal{M}$. In optimization, we compute a minimum of f; typical methods for solving this minimization problem are *iterative algorithms* on manifold $\mathcal{M}$. In an iterative algorithm based on line search, with a given starting point $w_0 \in \mathcal{M}$, we generate a sequence $\{w_t\}_{t\geq 0}$ on $\mathcal{M}$ that converges to w^* whenever w_0 is in a neighborhood of w^*. In an iterative optimization algorithm, we compute a search direction and then move in the search direction. More specifically, an iteration on manifold $\mathcal{M}$ is performed by following geodesics emanating from w_t and tangent to ξ_{w_t} at w_t. The notion of *geodesics* on Riemannian manifolds is a generalized concept of straight lines in Euclidean space. For any tangent vector $\xi \in T_w\mathcal{M}$ at $w \in \mathcal{M}$, there exist an interval I about 0 and a unique geodesic $\gamma_e(\cdot; w, \xi)\colon I \to \mathcal{M}$ such that $\gamma_e(0; w, \xi) = w$ and $\dot{\gamma}_e(0; w, \xi) = \xi$. The *exponential mapping* $\mathrm{Exp}_w\colon T_w\mathcal{M} \to \mathcal{M}$ at $w \in \mathcal{M}$ is defined by geodesics emanating

from w as $\mathrm{Exp}_w\xi = \gamma_e(1; w, \xi)$ for $\xi \in T_w\mathcal{M}$. If $\mathcal{M}$ is a complete manifold, the exponential mapping is defined for all vectors $\xi \in T_w\mathcal{M}$ [1, section 5.4]. We can thus obtain an update formula using the exponential mapping

$$w_{t+1} = \mathrm{Exp}_{w_t}(s_t\xi_{w_t}),$$

where the search direction ξ_{w_t} is in the tangent space $T_{w_t}\mathcal{M}$ of $\mathcal{M}$ at w_t, the scalar $s_t > 0$ is the step size, and $\mathrm{Exp}_{w_t}(\cdot)$ is the exponential mapping, which induces a line search algorithm along the geodesics. In addition, given two points w and z on $\mathcal{M}$, the *logarithm mapping*, or simply *log mapping*, which is the inverse of the exponential mapping, maps z to a vector $\xi \in T_w\mathcal{M}$ on the tangent space at w. The log mapping satisfies $\mathrm{dist}(w, z) = \|\mathrm{Log}_w(z)\|_w$, where $\mathrm{dist}\colon \mathcal{M} \times \mathcal{M} \to \mathbb{R}$ is the shortest distance between two points on $\mathcal{M}$ and $\|\cdot\|_w$ is the norm in $T_w\mathcal{M}$ defined through a Riemannian metric (see below).

The *steepest descent* or *full gradient descent* method for minimizing f on $\mathcal{M}$ is an iterative algorithm obtained when $-\mathrm{grad}f(w_t)$ is used as the search direction ξ_{w_t}. $\mathrm{grad}f(w_t)$ is the *Riemannian gradient* of f at w_t, which is computed according to the chosen metric g at $w_t \in \mathcal{M}$. Collecting each metric $g_w\colon T_w\mathcal{M} \times T_w\mathcal{M} \to \mathbb{R}$ over $w \in \mathcal{M}$ gives a family called a *Riemannian metric* on $\mathcal{M}$. $g_w(\xi_w, \zeta_w)$ is an inner product of elements ξ_w and ζ_w in the tangent space $T_w\mathcal{M}$ at w. Here, we use the notation $\langle\cdot,\cdot\rangle_w$ instead of $g(\cdot,\cdot)_w$ for simplicity. The gradient $\mathrm{grad}f(w)$ is defined as the unique element of $T_w\mathcal{M}$ that satisfies

$$\mathrm{D}f(w)[\xi_w] = \langle \mathrm{grad}f(w), \xi_w\rangle_w \quad \forall \xi_w \in T_w\mathcal{M},$$

where $\mathrm{D}f(w)\colon T_w\mathcal{M} \to \mathbb{R}$ is the *derivative* of f at w.

In searching for the next point along a geodesic, we need to compute tangent vectors obtained by the exponential mapping, which are expensive to compute in general. There are some Riemannian manifolds for which a closed form for the exponential mapping is not available. Alternatively, we can use curves other than geodesics as long as a starting point and its tangent vector at the initial time are the same as those of the geodesics. A more general update formula is then written as

$$w_{t+1} = R_{w_t}(s_t\xi_{w_t}),$$

where R_{w_t} is a *retraction*, which is any map $R_w\colon T_w\mathcal{M} \to \mathcal{M}$ that locally approximates the exponential mapping, up to the first order, on the manifold. The definition of a retraction is as follows [1, Definition 4.1.1].

DEFINITION 2.1. *$R\colon T\mathcal{M} \to \mathcal{M}$ is called a retraction on $\mathcal{M}$ if the restriction $R_w\colon T_w\mathcal{M} \to \mathcal{M}$ to $T_w\mathcal{M}$ for any $w \in \mathcal{M}$ satisfies both of the following:*

1. *$R_w(0_w) = w$, where 0_w is the zero vector in $T_w\mathcal{M}$;*
2. *$\mathrm{D}R_w(0_w)[\xi] = \xi$ for any $\xi \in T_w\mathcal{M}$.*

Retractions include the exponential mapping as a special case. An advantage of using retractions is that the computational cost can be reduced compared to exponential mapping. It is worth noting that the convergence properties for the exponential mapping usually hold for retractions as well.

In the R-SVRG proposed in section 3, we need to add tangent vectors that are in different tangent spaces, say $\tilde{w}$ and w on $\mathcal{M}$. A mathematically natural way to do so is to use the parallel translation operator. Parallel translation P_γ transports a vector field ξ on the geodesic curve γ that satisfies $P_\gamma^{a\leftarrow a}(\xi(a)) = \xi(a)$ and $\frac{\mathrm{D}}{\mathrm{d}t}(P_\gamma^{t\leftarrow a}\xi(a)) =$

0 [1, section 5.4], where $P_\gamma^{b\leftarrow a}$ is the parallel translation operator sending $\xi(a)$ to $\xi(b)$. However, parallel translation is sometimes computationally expensive, and no explicit formula is available for some manifolds, such as the Stiefel manifold. A *vector transport* on $\mathcal{M}$, which is a map $\mathcal{T}: T\mathcal{M} \oplus T\mathcal{M} \to T\mathcal{M}$, is used as an alternative. The definition of vector transport is as follows [1, Definition 8.1.1].

DEFINITION 2.2. *$\mathcal{T}: T\mathcal{M} \oplus T\mathcal{M} \to T\mathcal{M}$ is called a vector transport on $\mathcal{M}$ if it satisfies all of the following:*

1. *$\mathcal{T}$ has an associated retraction R, i.e., for $w \in \mathcal{M}$ and $\xi, \eta \in T_w\mathcal{M}$, $\mathcal{T}_\eta(\xi)$ is a tangent vector at $R_w(\xi)$;*
2. *$\mathcal{T}_{0_w}(\xi) = \xi$, where $\xi \in T_w\mathcal{M}$ and $w \in \mathcal{M}$;*
3. *$\mathcal{T}_\eta(a\xi + b\zeta) = a\mathcal{T}_\eta(\xi) + b\mathcal{T}_\eta(\zeta)$, where $a, b \in \mathbb{R}$, $\eta, \xi, \zeta \in T_w\mathcal{M}$, and $w \in \mathcal{M}$.*

With a vector transport $\mathcal{T}$, $\mathcal{T}_\eta(\xi)$ can be regarded as a transported vector of ξ along η. Parallel translation is an example of vector transport. In the following, we use the notation P_η and $P_\gamma^{z\leftarrow w}$ interchangeably, where γ is a curve connecting w and z on $\mathcal{M}$ defined by retraction R as $\gamma(\tau) := R_w(\tau\eta)$ with $z = R_w(\eta)$. We also omit the subscript γ when there is no confusion about the curve along which we transport a vector.

3. Riemannian stochastic variance reduced gradient algorithm. After a brief explanation of the variance reduced gradient variants in Euclidean space, we describe the proposed R-SVRG algorithm on Riemannian manifolds.

3.1. Variance reduced gradient variants in Euclidean space. The SGD update in Euclidean space is $w_{t+1} = w_t - \alpha v_t$, where v_t is a randomly selected vector called the *stochastic gradient* and α is the step size. SGD assumes an *unbiased estimator* of the full gradient as $\mathbb{E}_n[\nabla f_n(w_t)] = \nabla f(w_t)$. Many recent variants of the variance reduced gradient of SGD attempt to reduce its variance $\mathbb{E}[\|v_t - \nabla f(w_t)\|^2]$ as t increases to achieve better convergence [7, 15, 21, 27, 29, 30, 32]. SVRG, proposed in [15], introduces an explicit variance reduction strategy with double loops, where the sth outer loop, called the sth *epoch*, has m_s inner iterations. SVRG first keeps $\tilde{w}^{s-1} = w^{s-1}_{m_{s-1}}$ or $\tilde{w}^{s-1} = w^{s-1}_t$ for randomly chosen $t \in \{1, 2, \ldots, m_{s-1}\}$ at the end of the $(s-1)$th epoch and also sets the initial value of the sth epoch to $w^s_0 = \tilde{w}^{s-1}$. It then computes a full gradient $\nabla f(\tilde{w}^{s-1})$. Subsequently, denoting the selected random index $i \in \{1, 2, \ldots, N\}$ by i^s_t, SVRG randomly picks the i^s_tth sample for each $t \geq 1$ at $s \geq 1$ and computes the *modified stochastic gradient* v^s_t as

$$v^s_t = \nabla f_{i^s_t}(w^s_{t-1}) - \nabla f_{i^s_t}(\tilde{w}^{s-1}) + \nabla f(\tilde{w}^{s-1}). \tag{2}$$

Note that SVRG can be regarded as a special case of S2GD (semistochastic gradient descent), which differs in terms of the number of inner loop iterations chosen [19].

3.2. Proposed Riemannian extension of SVRG (R-SVRG). We propose a Riemannian extension of SVRG on a Riemannian manifold $\mathcal{M}$, called R-SVRG. Here, we denote the Riemannian stochastic gradient for the i^s_tth sample by $\mathrm{grad} f_{i^s_t}$ and the *modified Riemannian stochastic gradient* by ξ^s_t instead of v^s_t, in order to indicate the differences from the Euclidean case.

R-SVRG reduces the variance of the Riemannian stochastic gradient analogously to the SVRG algorithm in the Euclidean case. More specifically, R-SVRG keeps $\tilde{w}^{s-1} \in \mathcal{M}$ after m_{s-1} stochastic update steps of the $(s-1)$th epoch, and computes the full Riemannian gradient $\mathrm{grad} f(\tilde{w}^{s-1}) = \frac{1}{N}\sum_{i=1}^N \mathrm{grad} f_i(\tilde{w}^{s-1})$ only for this stored $\tilde{w}^{s-1}$. The algorithm also computes the Riemannian stochastic gradient $\mathrm{grad} f_{i^s_t}(\tilde{w}^{s-1})$

that corresponds to the i_t^sth sample. Picking the i_t^sth sample for each tth inner iteration of the sth epoch at w_{t-1}^s, we calculate ξ_t^s in the same way as v_t^s in (2), i.e., by modifying the stochastic gradient $\mathrm{grad} f_{i_t^s}(w_{t-1}^s)$ using both $\mathrm{grad} f(\tilde{w}^{s-1})$ and $\mathrm{grad} f_{i_t^s}(\tilde{w}^{s-1})$. Translating the right-hand side of (2) to manifold $\mathcal{M}$ involves the sum of $\mathrm{grad} f_{i_t^s}(w_{t-1}^s)$, $-\mathrm{grad} f_{i_t^s}(\tilde{w}^{s-1})$, and $\mathrm{grad} f(\tilde{w}^{s-1})$, which belong to two separate tangent spaces $T_{w_{t-1}^s}\mathcal{M}$ and $T_{\tilde{w}^{s-1}}\mathcal{M}$. This operation requires particular attention on a manifold, and a vector transport enables us to handle multiple elements on two separate tangent spaces flexibly. More specifically, $\mathrm{grad} f_{i_t^s}(\tilde{w}^{s-1})$ and $\mathrm{grad} f(\tilde{w}^{s-1})$ are first transported to $T_{w_{t-1}^s}\mathcal{M}$ at the current point, w_{t-1}^s; then, they can be added to $\mathrm{grad} f_{i_t^s}(w_{t-1}^s)$ on $T_{w_{t-1}^s}\mathcal{M}$. Consequently, the modified Riemannian stochastic gradient ξ_t^s at the tth inner iteration of the sth epoch is set to

$$\xi_t^s = \mathrm{grad} f_{i_t^s}(w_{t-1}^s) - \mathcal{T}_{\tilde{\eta}_{t-1}^s}(\mathrm{grad} f_{i_t^s}(\tilde{w}^{s-1}) - \mathrm{grad} f(\tilde{w}^{s-1})), \tag{3}$$

where $\tilde{\eta}_{t-1}^s \in T_{\tilde{w}^{s-1}}\mathcal{M}$ satisfies $R_{\tilde{w}^{s-1}}(\tilde{\eta}_{t-1}^s) = w_{t-1}^s$ and $\mathcal{T}$ is a vector transport operator on $\mathcal{M}$. Specifically, we need to calculate the tangent vector from $\tilde{w}^{s-1}$ to w_{t-1}^s, which is given by the inverse of the retraction, i.e., R^{-1}, if available. Consequently, the final update rule of R-SVRG is defined as $w_t^s = R_{w_{t-1}^s}(-\alpha_{t-1}^s \xi_t^s)$, where $\alpha_{t-1}^s > 0$ is the step size at w_{t-1}^s. As shown in our local convergence analysis (section 5.2), α_{t-1}^s can be fixed once the iterate becomes sufficiently close to a solution.

Let $\mathbb{E}_{i_t^s}[\cdot]$ be the expectation with respect to the random choice of i_t^s, conditioned on all the randomness introduced up to the tth iteration of the inner loop of the sth epoch. Conditioned on w_{t-1}^s, we take the expectation with respect to i_t^s and obtain

$$\begin{aligned}\mathbb{E}_{i_t^s}[\xi_t^s] &= \mathbb{E}_{i_t^s}[\mathrm{grad} f_{i_t^s}(w_{t-1}^s)] - \mathcal{T}_{\tilde{\eta}_{t-1}^s}(\mathbb{E}_{i_t^s}[\mathrm{grad} f_{i_t^s}(\tilde{w}^{s-1})] - \mathrm{grad} f(\tilde{w}^{s-1})) \\ &= \mathrm{grad} f(w_{t-1}^s) - \mathcal{T}_{\tilde{\eta}_{t-1}^s}(\mathrm{grad} f(\tilde{w}^{s-1}) - \mathrm{grad} f(\tilde{w}^{s-1})) \\ &= \mathrm{grad} f(w_{t-1}^s).\end{aligned}$$

The theoretical convergence analysis of the Euclidean SVRG algorithm assumes that the initial vector w_0^s of the sth epoch is set to the average (or a random) vector of the $(s-1)$th epoch [15, Figure 1]. However, the set of the last vectors in the $(s-1)$th epoch, i.e., $w_{m_{s-1}}^{s-1}$, gives a superior performance in the Euclidean SVRG algorithm. Therefore, for our local convergence rate analysis in Theorem 5.14, this study also uses, as option I, the mean value of $\tilde{w}^s = g_{m_s}(w_1^s, w_2^s, \ldots, w_{m_s}^s)$ as $\tilde{w}^s$, where $g_n(w_1, w_2, \ldots, w_n)$ is the Riemannian centroid on the manifold. Alternatively, we can also simply choose $\tilde{w}^s = w_t^s$ for $t \in \{1, 2, \ldots, m_s\}$ at random. In addition, as option II, we can use the last iterate in the sth epoch, i.e., $\tilde{w}^s = w_{m_s}^s$. Note that option II is always of practical use because no additional computation is needed to obtain $\tilde{w}^s$. However, if we can compute the Riemannian centroid of $w_1^s, w_2^s, \ldots, w_{m_s}^s$ relatively cheaply, option I may attain a better linear convergence rate. In the global convergence analysis in section 4 we use option II, whereas in the local convergence analysis in section 5 both options are analyzed. The overall algorithm is summarized in Algorithm 1.

In addition, variants of the variance reduced SGD initially require a full gradient calculation at every epoch. This initially results in a much greater overhead than the ordinary SGD algorithm and eventually induces a *cold-start* property in these variants. To avoid this, the use of the standard SGD update has been proposed only for the first epoch in Euclidean space [19]. We also adopt this simple modification of R-SVRG, denoted R-SVRG+; we do not analyze this extension but leave it as an open problem.

Algorithm 1 R-SVRG for problem (1).

Require: Update frequency $m_s > 0$ and sequence $\{\alpha_t^s\}$ of positive step sizes.

Initialize $\tilde{w}^0$.
for $s = 1, 2, \ldots$ **do**
  Calculate the full Riemannian gradient $\mathrm{grad} f(\tilde{w}^{s-1})$.
  Store $w_0^s = \tilde{w}^{s-1}$.
  for $t = 1, 2, \ldots, m_s$ **do**
    Choose $i_t^s \in \{1, 2, \ldots, N\}$ uniformly at random.
    Calculate the tangent vector $\tilde{\eta}_{t-1}^s$ from $\tilde{w}^{s-1}$ to w_{t-1}^s satisfying $R_{\tilde{w}^{s-1}}(\tilde{\eta}_{t-1}^s) = w_{t-1}^s$.
    Calculate the modified Riemannian stochastic gradient ξ_t^s by transporting $\mathrm{grad} f(\tilde{w}^{s-1})$ and $\mathrm{grad} f_{i_t^s}(\tilde{w}^{s-1})$ along $\tilde{\eta}_{t-1}^s$ as

$$\xi_t^s = \mathrm{grad} f_{i_t^s}(w_{t-1}^s) - \mathcal{T}_{\tilde{\eta}_{t-1}^s}(\mathrm{grad} f_{i_t^s}(\tilde{w}^{s-1}) - \mathrm{grad} f(\tilde{w}^{s-1})).$$

    Update w_t^s from w_{t-1}^s as $w_t^s = R_{w_{t-1}^s}(-\alpha_{t-1}^s \xi_t^s)$ with retraction R.
  end for
  Option I: $\tilde{w}^s = g_{m_s}(w_1^s, w_2^s, \ldots, w_{m_s}^s)$.
  Option II: $\tilde{w}^s = w_{m_s}^s$.
end for

As mentioned earlier, each iteration of R-SVRG has double loops to reduce the variance of the Riemannian stochastic gradient $\mathrm{grad} f_{i_t^s}(\tilde{w}^{s-1})$. The sth epoch, i.e., the outer loop, requires $N + 2m_s$ gradient evaluations, where N is for the full gradient $\mathrm{grad} f(\tilde{w}^{s-1})$ at the beginning of each sth epoch and $2m_s$ is for the inner iterations, since each inner step needs two gradient evaluations, i.e., $\mathrm{grad} f_{i_t^s}(w_{t-1}^s)$ and $\mathrm{grad} f_{i_t^s}(\tilde{w}^{s-1})$. However, if $\mathrm{grad} f_{i_t^s}(\tilde{w}^{s-1})$ for each sample is stored at the beginning of the sth epoch, as in SAG, the evaluations over all the inner loops result in m_s. Finally, the sth epoch requires $N + m_s$ evaluations. It is natural to choose an m_s of the same order as N but slightly larger (e.g., $m_s = 5N$ for nonconvex problems has been suggested in [15]).

4. Global convergence analysis. In this section, we present a global convergence analysis of Algorithm 1 for problem (1) after introducing some assumptions. Throughout this section, we let R and $\mathcal{T}$ denote a retraction and vector transport used in Algorithm 1, respectively, and we make the following assumptions.

Assumption 4.1. The retraction R is such that $R_w \colon T_w\mathcal{M} \to \mathcal{M}$ for any $w \in \mathcal{M}$ is of class C^2. The vector transport $\mathcal{T}$ is continuous and isometric on $\mathcal{M}$, i.e., for any $w \in \mathcal{M}$ and $\xi, \eta \in T_w\mathcal{M}$, $\|\mathcal{T}_\eta(\xi)\|_{R_w(\eta)} = \|\xi\|_w$.

Note that we can construct an isometric vector transport as in [11, 12] such that Assumption 4.1 holds.

Assumption 4.2. The objective function f is thrice continuously differentiable and its components $f_1, f_2, \ldots, f_N$ are twice continuously differentiable.

Assumption 4.3. For a sequence $\{w_t^s\}$ generated by Algorithm 1, there exists a compact and connected set $K \subset \mathcal{M}$ such that $w_t^s \in K$ for all $s, t \geq 0$. In addition, for each $s \geq 1$ and $t \geq 0$, there exists $\tilde{\eta}_{t-1}^s \in T_{\tilde{w}^{s-1}}\mathcal{M}$ such that $R_{\tilde{w}^{s-1}}(\tilde{\eta}_{t-1}^s) = w_{t-1}^s$.

Furthermore, there exists $I > 0$ such that, for any $z \in K$, $R_z(\cdot)$ is defined in a ball $\mathbb{B}(0_z, I) \subset T_z\mathcal{M}$, which is centered at the origin 0_z in $T_z\mathcal{M}$ with radius I.

Note that the existence of $\tilde{\eta}_{t-1}^s$ in Assumption 4.3 is guaranteed if $R_z(\cdot)$ for any $z \in K$ is a diffeomorphism in a ball $\mathbb{B}(0_z, \rho)$ that satisfies $K \subset R_z(\mathbb{B}(0_z, \rho))$. This assumption is a weakened version of the assumptions in some existing studies for special cases. For example, in [34], where the exponential mapping Exp is used as a retraction, the inverse of Exp is assumed to exist at all points $\{w_t^s\}$ in their global convergence analysis. Here, we note that if I in Assumption 4.3 is too small, then $R_{\tilde{w}^{s-1}}^{-1}$ may not be defined in a neighborhood of w_{t-1}^s. We just suppose that there exists a $\tilde{\eta}_{t-1}^s$ that is mapped to w_{t-1}^s by $R_{\tilde{w}^{s-1}}$. Further analysis with respect to the specific manifold in question may be required to specify the value of I.

In the global convergence analysis, we also assume that the sequence of step sizes $\{\alpha_t^s\}_{t\geq 0, s\geq 1}$ satisfies the usual condition in stochastic approximation as follows.

Assumption 4.4. The sequence $\{\alpha_t^s\}$ of step sizes satisfies

$$\sum (\alpha_t^s)^2 < \infty \quad \text{and} \quad \sum \alpha_t^s = \infty, \tag{4}$$

where $\sum$ denotes $\sum_{s=1}^{\infty} \sum_{t=0}^{m_s}$.

This condition is satisfied if, for example, $\{m_s\}$ is upper bounded and $\alpha_t^s = \alpha_0(1 + \alpha_0\lambda\lfloor k/m_s \rfloor)^{-1}$ with positive constants α_0 and λ, where k is the total iteration number depending on s and t and $\lfloor \cdot \rfloor$ denotes the floor function. We also note the following proposition introduced in [8].

PROPOSITION 4.5 (see [8]). *Let $(X_n)_{n\in\mathbb{N}}$ be a nonnegative stochastic process with bounded positive variations, i.e., such that $\sum_{n=0}^{\infty} \mathbb{E}[\mathbb{E}[X_{n+1} - X_n \mid \mathcal{F}_n]^+] < \infty$, where X^+ denotes the quantity $\max\{X, 0\}$ for a random variable X and $\mathcal{F}_n$ is the increasing sequence of σ-algebras generated by the variables just before time n. Then, the process is a quasi martingale, i.e.,*

$$\sum_{n=0}^{\infty} |\mathbb{E}[X_{n+1} - X_n \mid \mathcal{F}_n]| < \infty \text{ a.s.} \quad \text{and} \quad X_n \text{ converges a.s.}$$

Now, we give the almost sure convergence of the proposed algorithm under the assumption that the generated sequence is in a compact set.

THEOREM 4.6. *Suppose Assumptions* 4.1–4.3 *hold, and consider Algorithm* 1 *with option* II *and step sizes $\{\alpha_t^s\}$ satisfying Assumption* 4.4 *on a Riemannian manifold $\mathcal{M}$. If $f \geq 0$, then $\{f(w_t^s)\}$ converges a.s. and $\operatorname{grad} f(w_t^s) \to 0$ a.s.*

Proof. The claim is proved similarly to the proof of the standard Riemannian SGD (see [5]). Since K is compact, all continuous functions on K can be bounded. Therefore, there exists $C > 0$ such that for all $w \in K$ and $n \in \{1, 2, \ldots, N\}$ we have $\|\operatorname{grad} f(w)\|_w \leq C$ and $\|\operatorname{grad} f_n(w)\|_w \leq C$. We use $C' := 3C$ in the following. Moreover, as $\alpha_t^s \to 0$, there exists s_0 such that for $s \geq s_0$, we have $\alpha_t^s < 1$ and $\alpha_t^s C' < I$. Now, suppose that $s \geq s_0$. Let $\tilde{\eta}_t^s$ satisfy $R_{\tilde{w}^{s-1}}(\tilde{\eta}_t^s) = w_t^s$. The existence of such $\tilde{\eta}_t^s$ is guaranteed from Assumption 4.3. It follows from the triangle inequality that

$$\begin{aligned}\|\xi_{t+1}^s\|_{w_t^s} &= \|\operatorname{grad} f_{i_{t+1}^s}(w_t^s) - \mathcal{T}_{\tilde{\eta}_t^s}(\operatorname{grad} f_{i_{t+1}^s}(\tilde{w}^{s-1})) + \mathcal{T}_{\tilde{\eta}_t^s}(\operatorname{grad} f(\tilde{w}^{s-1}))\|_{w_t^s} \\ &\leq C + C + C = C',\end{aligned}$$

where we have used the assumption that $\mathcal{T}$ is an isometry. Therefore, we have

$$\|-\alpha_t^s \xi_{t+1}^s\|_{w_t^s} \leq \alpha_t^s C' < I.$$

Hence, it follows from Assumption 4.3 that there exists a curve $R_{w_t^s}(-\tau\alpha_t^s\xi_{t+1}^s)_{0\leq\tau\leq 1}$ linking w_t^s and w_{t+1}^s. Defining $g(\tau; w, \xi) := (f \circ R_w)(-\tau\xi)$, there exists a constant k_1 such that

$$\frac{d^2}{d\tau^2} g(\tau; w, \xi) \leq 2k_1, \qquad \tau \in [0,1],\ \xi \in \mathbb{B}(0_w, I),\ w \in K,$$

since $[0,1] \times \{(w,\xi) \mid \xi \in \mathbb{B}(0_w, I), w \in K\}$ is compact. A trivial relation

$$g(\alpha; w, \xi) = g(0; w, \xi) + g'(0; w, \xi)\alpha + \alpha^2 \int_0^1 (1-\tau) g''(\alpha\tau; w, \xi) d\tau,$$

together with $\alpha_t^s < 1$, then implies that

$$\begin{aligned} f(w_{t+1}^s) - f(w_t^s) &= f(R_{w_t^s}(-\alpha_t^s \xi_{t+1}^s)) - f(w_t^s) \\ &= g(\alpha_t^s; w_t^s, \xi_{t+1}^s) - g(0; w_t^s, \xi_{t+1}^s) \\ &= -\alpha_t^s \langle \xi_{t+1}^s, \mathrm{grad} f(w_t^s)\rangle_{w_t^s} + (\alpha_t^s)^2 \int_0^1 (1-\tau) g''(\alpha_t^s \tau) d\tau \\ &\leq -\alpha_t^s \langle \xi_{t+1}^s, \mathrm{grad} f(w_t^s)\rangle_{w_t^s} + (\alpha_t^s)^2 k_1. \end{aligned}$$

Let $\mathcal{F}_t^s$ be the increasing sequence of σ-algebras defined by

$$\mathcal{F}_t^s = \{i_1^1, i_2^1, \ldots, i_{m_1}^1, i_1^2, i_2^2, \ldots, i_{m_2}^2, \ldots, i_1^{s-1}, i_2^{s-1}, \ldots, i_{m_{s-1}}^{s-1}, i_1^s, i_2^s, \ldots, i_{t-1}^s\}.$$

Since w_t^s is computed from $i_1^1, i_2^1, \ldots, i_t^s$, it is measurable in $\mathcal{F}_{t+1}^s$. As i_{t+1}^s is independent of $\mathcal{F}_{t+1}^s$, we have

$$\begin{aligned} \mathbb{E}[\xi_{t+1}^s \mid \mathcal{F}_{t+1}^s] &= \mathbb{E}_{i_{t+1}^s}[\xi_{t+1}^s] \\ &= \mathbb{E}_{i_{t+1}^s}[\mathrm{grad} f_{i_{t+1}^s}(w_t^s) - \mathcal{T}_{\tilde{\eta}_t^s}(\mathrm{grad} f_{i_{t+1}^s}(\tilde{w}^{s-1})) + \mathcal{T}_{\tilde{\eta}_t^s}(\mathrm{grad} f(\tilde{w}^{s-1}))] \\ &= \mathrm{grad} f(w_t^s) \end{aligned}$$

from the linearity of $\mathcal{T}_{\tilde{\eta}_t^s}$. Therefore, it holds that

$$\mathbb{E}[\langle \xi_{t+1}^s, \mathrm{grad} f(w_t^s)\rangle_{w_t^s} \mid \mathcal{F}_{t+1}^s] = \|\mathrm{grad} f(w_t^s)\|_{w_t^s}^2,$$

which yields

$$\mathbb{E}[f(w_{t+1}^s) - f(w_t^s) \mid \mathcal{F}_{t+1}^s] \leq -\alpha_t^s \|\mathrm{grad} f(w_t^s)\|_{w_t^s}^2 + (\alpha_t^s)^2 k_1 \leq (\alpha_t^s)^2 k_1. \tag{5}$$

We reindex the sequence $\{w_0^1, w_1^1, \ldots, w_{m_1}^1 (= w_0^2), w_1^2, w_2^2, \ldots, w_{m_2}^2, \ldots, w_t^s, \ldots\}$ as $\{w_1, w_2, \ldots, w_k, \ldots\}$. We also similarly reindex $\{\alpha_t^s\}$. As $f(w_k) \geq 0$, (5) proves that $\{f(w_k) + \sum_{l=k}^{\infty} (\alpha_l)^2 k_1\}$ is a nonnegative supermartingale. Hence, it converges a.s. Returning to the original indexing, this implies that $\{f(w_t^s)\}$ converges a.s. Moreover, summing the inequalities, we have

$$\sum_{s\geq s_0, t} \alpha_t^s \|\mathrm{grad} f(w_t^s)\|_{w_t^s}^2 \leq -\sum_{s\geq s_0, t} \mathbb{E}[f(w_{t+1}^s) - f(w_t^s) \mid \mathcal{F}_{t+1}^s] + \sum_{s\geq s_0, t} (\alpha_t^s)^2 k_1. \tag{6}$$

Here, we prove that the right-hand side is bounded. By doing so, we can conclude the convergence of the left-hand term.

Summing (5) over s and t, we have

$$\sum_{s\geq s_0,t} \mathbb{E}[\mathbb{E}[f(w_{t+1}^s) - f(w_t^s) \mid \mathcal{F}_{t+1}^s]^+] < \infty$$

by assumption (4). It then follows from Proposition 4.5 that

$$\left| -\sum_{s\geq s_0,t} \mathbb{E}[f(w_{t+1}^s) - f(w_t^s) \mid \mathcal{F}_{t+1}^s] \right| \leq \sum_{s\geq s_0,t} |\mathbb{E}[f(w_{t+1}^s) - f(w_t^s) \mid \mathcal{F}_{t+1}^s]| < \infty,$$

which, together with (6), implies that $\sum_{s\geq s_0,t} \alpha_t^s \|\mathrm{grad} f(w_t^s)\|_{w_t^s}^2$ converges a.s. If we further prove that $\{\|\mathrm{grad} f(w_t^s)\|_{w_t^s}\}$ converges a.s., it can only converge to 0 a.s. because of (4).

As in [5], we consider the nonnegative process $p_t^s = \|\mathrm{grad} f(w_t^s)\|_{w_t^s}^2$. Bounding the largest eigenvalue of the Hessian of $\|\mathrm{grad} f(w)\|_w^2$ from above by k_2 on the compact set K along the curve defined by the retraction R linking w_t^s and w_{t+1}^s, a Taylor expansion yields

$$p_{t+1}^s - p_t^s \leq -2\alpha_t^s \langle \mathrm{grad} f(w_t^s), \mathrm{Hess} f(w_t^s)[\xi_{t+1}^s] \rangle_{w_t^s} + (\alpha_t^s)^2 \|\xi_{t+1}^s\|_{w_t^s}^2 k_2.$$

Let k_3 be a constant such that $-k_3$ is a lower bound of the minimum eigenvalue of the Hessian of f on K. Then, we have

$$\mathbb{E}[p_{t+1}^s - p_t^s \mid \mathcal{F}_{t+1}^s] \leq 2\alpha_t^s \|\mathrm{grad} f(w_t^s)\|_{w_t^s}^2 k_3 + (\alpha_t^s)^2 C'^2 k_2.$$

We have proved the fact that the infinite series of the terms on the right-hand side is finite, implying that $\{p_t^s\}$ is a quasi martingale, which further implies the a.s. convergence of $\{p_t^s\}$ to 0, as claimed. ∎

Theorem 4.6 includes a global convergence of the algorithm with exponential mapping and parallel translation as a special case. In this case, a sufficient condition for the assumptions can easily be written as the following corollary by using the notion of injectivity radius.

COROLLARY 4.7. *Suppose Assumption* 4.1 *holds and consider Algorithm* 1 *with option* II *and step sizes* $\{\alpha_t^s\}$ *satisfying Assumption* 4.4 *on a Riemannian manifold* $\mathcal{M}$*, where exponential mapping and parallel translation are used as retraction and vector transport, respectively. Assume that* $\mathcal{M}$ *is connected and has an injectivity radius uniformly bounded from below by a positive constant. Assume also that there exists a compact and connected set* $K \subset \mathcal{M}$ *such that* $w_t^s \in K$ *for all* $s, t \geq 0$*. If* $f \geq 0$*, then* $\{f(w_t^s)\}$ *converges a.s. and* $\mathrm{grad} f(w_t^s) \to 0$ *a.s.*

5. Local convergence rate analysis. In this section, we show the local convergence rate analysis of the R-SVRG algorithm; we analyze the convergence of any sequences generated by Algorithm 1 that are contained in a sufficiently small neighborhood of a local minimum point of the objective function. Hence, we can assume that the objective function is strongly convex in such a neighborhood. We first give some formal expressions of these assumptions and then analyze the local convergence rate of the algorithm.

5.1. Assumptions and existing lemmas. We again suppose Assumption 4.1 holds and make the following additional assumption, which is a weaker version of Assumption 4.2.

Assumption 5.1. The objective function f and its components $f_1, f_2, \ldots, f_N$ are twice continuously differentiable.

Let w^* be a critical point of f. As discussed in [12], for a positive constant ρ, a ρ-totally retractive neighborhood Ω of $w \in \mathcal{M}$ is a neighborhood such that, for all $z \in \Omega$, $\Omega \subset R_z(\mathbb{B}(0_z, \rho))$ and $R_z(\cdot)$ is a diffeomorphism on $\mathbb{B}(0_z, \rho)$, which is the ball in $T_z\mathcal{M}$ with center 0_z and radius ρ, where 0_z is the zero vector in $T_z\mathcal{M}$. The concept of a totally retractive neighborhood is analogous to that of a totally normal neighborhood for exponential mapping. Note that, for each point, a ρ-totally retractive neighborhood for sufficiently small $\rho > 0$ is proved to exist [12]. On the other hand, it should be noted that a concrete bound of the radii of locally retractive neighborhoods is difficult to specify. In the local convergence analysis, we assume the following.

Assumption 5.2. The sequence $\{w_t^s\}$ generated by Algorithm 1 remains continuously in the totally retractive neighborhood Ω of the critical point w^* of f, i.e., $R_{w_t^s}(-\alpha\xi_{t+1}^s) \in \Omega$ for all $s, t \geq 0$ and for all $\alpha \in [0, \alpha_t^s]$. In addition, the radius of Ω is sufficiently small.

We note that the assumption on the radius of Ω in Assumption 5.2 is for Assumption 5.4 and for Lemma 5.12.

Assumption 5.3. There exists a constant c_0 such that vector transport $\mathcal{T}$ satisfies the following conditions:

$$\|\mathcal{T}_\eta - \mathcal{T}_{R_\eta}\| \leq c_0\|\eta\|, \qquad \|\mathcal{T}_\eta^{-1} - \mathcal{T}_{R_\eta}^{-1}\| \leq c_0\|\eta\|,$$

where $\|\cdot\|$ denotes the induced (operator) norm of the Riemannian metric and $\mathcal{T}_R$ denotes the differentiated retraction, i.e.,

$$\mathcal{T}_{R_\eta}(\xi) = \mathrm{D}R_w(\eta)[\xi], \qquad \eta, \xi \in T_w\mathcal{M},\ w \in \mathcal{M}.$$

Assumption 5.3 states that the difference between $\mathcal{T}$ and $\mathcal{T}_R$ is small if the tangent vector along which a vector is transported is close to 0. See [12] for further details.

Furthermore, we assume the following.

Assumption 5.4. f is strongly retraction-convex with respect to R in Ω; i.e., there exist two constants $0 < a_0 < a_1$ such that $a_0 \leq \frac{d^2}{d\alpha^2} f(R_w(\alpha\eta)) \leq a_1$ for all $w \in \Omega$, $\eta \in T_w\mathcal{M}$ with $\|\eta\|_w = 1$, and for all α satisfying $R_w(\tau\eta) \in \Omega$ for all $\tau \in [0, \alpha]$. Furthermore, f is strongly geodesically convex in Ω, i.e., f is strongly retraction-convex with respect to Exp.

Letting Ω be smaller if necessary, we can guarantee the last assumption from the other assumptions using [12, Lemma 3.1].

The next assumption states how close the modified stochastic gradient by retraction R is to that obtained by the exponential mapping.

Assumption 5.5. Let $\{\xi_t^s\}$ be generated by Algorithm 1 with a fixed step size of $\alpha_t^s := \alpha$. Exp_w^{-1} for any $w \in \Omega$ is defined in Ω and there exists a constant $c_R > 0$ such that $\|\mathrm{Exp}_{w_{t-1}^s}^{-1}(w_t^s) - (-\alpha\xi_t^s)\|_{w_{t-1}^s} \leq c_R\|\alpha\xi_t^s\|_{w_{t-1}^s}^2$.

If $R^{-1}_{w^s_{t-1}}$ is defined in a ball whose image by $R_{w^s_{t-1}}$ contains Ω, then the above assumption translates into

$$\|\mathrm{Exp}^{-1}_{w^s_{t-1}}(w^s_t) - R^{-1}_{w^s_{t-1}}(w^s_t)\|_{w^s_{t-1}} \leq c_R \|R^{-1}_{w^s_{t-1}}(w^s_t)\|^2_{w^s_{t-1}},$$

which is natural when we assume sufficient smoothness of Exp and R because R coincides with Exp up to the first order.

In the remainder of this section, we introduce some existing lemmas to evaluate the differences between using retraction and vector transport and using exponential mapping and parallel translation, and the effects of the curvature of the manifold in question.

LEMMA 5.6 (see the proof of [12, Lemma 3.9]). *Under Assumptions* 5.1 *and* 5.2*, there exists a constant $\beta > 0$ such that*

$$\|P^{w \leftarrow z}_{\gamma}(\mathrm{grad} f(z)) - \mathrm{grad} f(w)\|_w \leq \beta \,\mathrm{dist}(z, w), \tag{7}$$

where w and z are in Ω and γ is a curve $\gamma(\tau) := R_z(\tau\eta)$ for an arbitrary $\eta \in T_z\mathcal{M}$ defined by retraction R on $\mathcal{M}$. $P^{w \leftarrow z}_{\gamma}(\cdot)$ is a parallel translation operator along curve γ from z to w.

Note that curve γ in this lemma is not necessarily the geodesic on $\mathcal{M}$. Relation (7) is a generalization of the Lipschitz continuity condition.

LEMMA 5.7 (see [12, Lemma 3.5]). *Let $\mathcal{T} \in C^0$ be a vector transport associated with the same retraction R as that of the parallel transport $P \in C^\infty$. Under Assumption* 5.3*, for any $\bar{w} \in \mathcal{M}$, there exists a constant $\theta > 0$ and neighborhood $\mathcal{U}$ of $\bar{w}$ such that for all $w, z \in \mathcal{U}$,*

$$\|\mathcal{T}_\eta(\xi) - P_\eta(\xi)\|_z \leq \theta \|\xi\|_w \|\eta\|_w,$$

where $\xi, \eta \in T_w\mathcal{M}$ and $R_w(\eta) = z$.

We can derive the following lemma from the Taylor expansion as in the proof of [12, Lemma 3.2].

LEMMA 5.8 (see the proof of [12, Lemma 3.2]). *Under Assumptions* 5.2–5.4*, there exists a positive real number σ such that*

$$f(z) \geq f(w) + \left\langle \mathrm{Exp}^{-1}_w(z), \mathrm{grad} f(w) \right\rangle_w + \frac{\sigma}{2} \left\| \mathrm{Exp}^{-1}_w(z) \right\|^2_w, \qquad w, z \in \Omega. \tag{8}$$

Proof. From the assumptions, there exists $\sigma > 0$ such that $\frac{d^2}{d\alpha^2} f(\mathrm{Exp}_w(\alpha\eta)) \geq \sigma$ for all $w \in \Omega$, $\eta \in T_w\mathcal{M}$ with $\|\eta\|_w = 1$, and for all α such that $\mathrm{Exp}_w(\tau\eta) \in \Omega$ for all $\tau \in [0, \alpha]$. From Taylor's theorem, we can conclude that this σ satisfies the claim. ∎

If we replace Exp in Lemma 5.8 with retraction R on $\mathcal{M}$, we can obtain a similar result for R, which is shown in [12, Lemma 3.2], where the constant a_0 corresponds to σ in our Lemma 5.8. However, Lemma 5.8 for Exp is sufficient in the following discussion.

From [11, Lemma 3], we have the following.

LEMMA 5.9 (see [11, Lemma 3]). *Let $\mathcal{M}$ be a Riemannian manifold endowed with retraction R and let $\bar{w} \in \mathcal{M}$. Then, there exist $\mu > 0$, $\nu > 0$, and $\delta_{\mu,\nu} > 0$ such that, for all w in a sufficiently small neighborhood of $\bar{w}$ and all $\xi \in T_w\mathcal{M}$ with $\|\xi\|_w \leq \delta_{\mu,\nu}$, the inequalities*

$$\|\xi\|_w \leq \mu \,\mathrm{dist}(w, R_w(\xi)) \quad \text{and} \quad \mathrm{dist}(w, R_w(\xi)) \leq \nu \|\xi\|_w \tag{9}$$

hold.

Since we have $\mathrm{dist}(w, \mathrm{Exp}_w(\xi)) = \|\xi\|_w$, (9) is equivalent to $\mathrm{dist}(w, \mathrm{Exp}_w(\xi)) \leq \mu\, \mathrm{dist}(w, R_w(\xi))$ and $\mathrm{dist}(w, R_w(\xi)) \leq \nu\, \mathrm{dist}(w, \mathrm{Exp}_w(\xi))$, which give relations between the exponential mapping and a general retraction.

Now, we introduce [35, Lemma 6] to evaluate the distance between w_t^s and w^* using the smoothness of our objective function. In the following, an Alexandrov space is defined as a length space whose curvature is bounded.

LEMMA 5.10 (see [35, Lemma 6]). *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*

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

5.2. Local convergence rate analysis with retraction and vector transport. We now demonstrate the local convergence properties of the R-SVRG algorithm (i.e., local convergence to local minimizers) and its convergence rate. The main theorem (Theorem 5.14) follows three lemmas, the proofs of which are provided in Appendix A.

The following lemma shows a property of the Riemannian centroid on a general Riemannian manifold.

LEMMA 5.11. *Let $w_1, w_2, \ldots, w_m$ be points on a Riemannian manifold $\mathcal{M}$ and let w be the Riemannian centroid of the m points. For an arbitrary point p on $\mathcal{M}$, we have*

$$(\mathrm{dist}(p, w))^2 \leq \frac{4}{m} \sum_{i=1}^{m} (\mathrm{dist}(p, w_i))^2.$$

Then, we state that the norms of modified stochastic gradients in R-SVRG are sufficiently small under Assumptions 4.1 and 5.1–5.3.

LEMMA 5.12. *Under Assumptions* 4.1 *and* 5.1–5.3*, the norm of ξ_t^s computed by* (3) *is sufficiently small; i.e., for any $\varepsilon > 0$, there exists $r_0 > 0$ such that $\|\xi_t^s\|_{w_{t-1}^s} \leq \varepsilon$ for r satisfying $r < r_0$, where r is the radius of Ω.*

We now provide the upper bound of the variance of ξ_t^s as follows.

LEMMA 5.13. *Suppose Assumptions* 4.1 *and* 5.1–5.3 *hold, which guarantees Lemmas* 5.6, 5.7, *and* 5.9 *for $\bar{w} = w^*$. Let $\beta > 0$ be a constant such that*

$$\|P_\gamma^{w \leftarrow z}(\mathrm{grad} f_n(z)) - \mathrm{grad} f_n(w)\|_w \leq \beta\, \mathrm{dist}(z, w), \qquad w, z \in \Omega,\ n = 1, 2, \ldots, N.$$

The existence of such a β is guaranteed by Lemma 5.6. *The upper bound of the variance of ξ_t^s is given by*

$$\mathbb{E}_{i_t^s}[\|\xi_t^s\|_{w_{t-1}^s}^2] \leq 4(\beta^2 + \mu^2 C^2 \theta^2)(7(\mathrm{dist}(w_{t-1}^s, w^*))^2 + 4(\mathrm{dist}(\tilde{w}^{s-1}, w^*))^2), \tag{10}$$

where the constant θ corresponds to that in Lemma 5.7*, C is the upper bound of $\|\mathrm{grad} f_n(w)\|$, $n = 1, 2, \ldots, N$, for $w \in \Omega$, and $\mu > 0$ appears in* (9).

We proceed to the main theorem and prove it using Lemmas 5.11–5.13.

THEOREM 5.14. *Let $\mathcal{M}$ be a Riemannian manifold with curvature lower bounded by κ, $w^* \in \mathcal{M}$ be a nondegenerate local minimizer of f (i.e.,* $\mathrm{grad} f(w^*) = 0$ *and the Hessian* $\mathrm{Hess} f(w^*)$ *of f at w^* is positive definite), D be the diameter of the compact set Ω, and $\zeta := \sqrt{|\kappa|}D/\tanh(\sqrt{|\kappa|}D)$ if $\kappa < 0$ and $\zeta := 1$ if $\kappa \geq 0$. Suppose Assumptions* 4.1 *and* 5.1–5.5 *hold. Let c_R be the constant in Assumption* 5.5*, let β, μ, θ, and C be the same as in Lemma* 5.13*, and let ν be the constant in Lemma* 5.9*. Suppose that $\sigma > 0$ is the constant in Lemma* 5.8 *satisfying* (8)*. Let α be a positive number satisfying $0 < \alpha(\sigma - 28(\zeta\nu + 2c_R D)(\beta^2 + \mu^2 C^2\theta^2)\alpha) < 1$. It then follows that, for any sequence $\{\tilde{w}^s\}$ generated by Algorithm* 1 *with a fixed step size $\alpha_t^s := \alpha$ and $m_s := m$ converging to w^*, there exists $K > 0$ such that, for all $s > K$,*

$$\mathbb{E}[(\mathrm{dist}(\tilde{w}^s, w^*))^2] \leq \delta \mathbb{E}[(\mathrm{dist}(\tilde{w}^{s-1}, w^*))^2], \tag{11}$$

where

$$\delta := \begin{cases} \dfrac{4(1 + 16m(\zeta\nu + 2c_R D)(\beta^2 + \mu^2 C^2\theta^2)\alpha^2)}{m\alpha(\sigma - 28(\zeta\nu + 2c_R D)(\beta^2 + \mu^2 C^2\theta^2)\alpha)} & \text{with option I,} \\ 1 - \sigma\alpha + 4(4m+7)(\zeta\nu + 2c_R D)(\beta^2 + \mu^2 C^2\theta^2)\alpha^2 & \text{with option II.} \end{cases}$$

Before proving the theorem, we summarize the above theorem as the following corollary.

COROLLARY 5.15. *Let $\mathcal{M}$ be a Riemannian manifold whose curvature is lower bounded and let $w^* \in \mathcal{M}$ be a nondegenerate local minimizer of f. Suppose Assumptions* 4.1 *and* 5.1–5.4 *hold, and let α be a positive number. If α is sufficiently small, for any sequence $\{\tilde{w}^s\}$ generated by Algorithm* 1 *with a fixed step size $\alpha_t^s := \alpha$ and $m_s := m$ converging to w^*, there exists $K > 0$ such that, for all $s > K$,*

$$\mathbb{E}[(\mathrm{dist}(\tilde{w}^s, w^*))^2] \leq \delta \mathbb{E}[(\mathrm{dist}(\tilde{w}^{s-1}, w^*))^2],$$

where δ is a constant in $(0, 1)$.

Proof of Theorem 5.14. Since the function $x/\tanh(x)$ on x monotonically increases in $[0, \infty)$, we have, from Lemma 5.10,

$$a^2 \leq \zeta b^2 + c^2 - 2bc\cos(A)$$

for any geodesic triangle in Ω with side lengths a, b, and c, since $\sqrt{|\kappa|}c \leq \sqrt{|\kappa|}D$. Then, conditioned on w_{t-1}^s, the expectation of the distance between w_t^s and w^* with respect to the random choice of i_t^s is evaluated by considering the geodesic triangle with w_{t-1}^s, w^*, and w_t^s in Ω as

$$\mathbb{E}_{i_t^s}[(\mathrm{dist}(w_t^s, w^*))^2] \leq \mathbb{E}_{i_t^s}\Big[\zeta(\mathrm{dist}(w_{t-1}^s, w_t^s))^2 + (\mathrm{dist}(w_{t-1}^s, w^*))^2 - 2\big\langle \mathrm{Exp}_{w_{t-1}^s}^{-1}(w_t^s), \mathrm{Exp}_{w_{t-1}^s}^{-1}(w^*)\big\rangle_{w_{t-1}^s}\Big].$$

From Assumption 5.5,

$$\begin{aligned} &-\big\langle \mathrm{Exp}_{w_{t-1}^s}^{-1}(w_t^s), \mathrm{Exp}_{w_{t-1}^s}^{-1}(w^*)\big\rangle_{w_{t-1}^s} \\ &\quad = \big\langle -\alpha\xi_t^s - \mathrm{Exp}_{w_{t-1}^s}^{-1}(w_t^s), \mathrm{Exp}_{w_{t-1}^s}^{-1}(w^*)\big\rangle_{w_{t-1}^s} - \big\langle -\alpha\xi_t^s, \mathrm{Exp}_{w_{t-1}^s}^{-1}(w^*)\big\rangle_{w_{t-1}^s} \\ &\quad \leq \|-\alpha\xi_t^s - \mathrm{Exp}_{w_{t-1}^s}^{-1}(w_t^s)\|_{w_{t-1}^s} \|\mathrm{Exp}_{w_{t-1}^s}^{-1}(w^*)\|_{w_{t-1}^s} - \big\langle -\alpha\xi_t^s, \mathrm{Exp}_{w_{t-1}^s}^{-1}(w^*)\big\rangle_{w_{t-1}^s} \end{aligned}$$

$$\leq c_R\|\alpha\xi_t^s\|_{w_{t-1}^s}^2\|\mathrm{Exp}_{w_{t-1}^s}^{-1}(w^*)\|_{w_{t-1}^s} - \left\langle -\alpha\xi_t^s, \mathrm{Exp}_{w_{t-1}^s}^{-1}(w^*)\right\rangle_{w_{t-1}^s}$$
$$\leq c_R D\|\alpha\xi_t^s\|_{w_{t-1}^s}^2 - \left\langle -\alpha\xi_t^s, \mathrm{Exp}_{w_{t-1}^s}^{-1}(w^*)\right\rangle_{w_{t-1}^s},$$

in which the relation $\|\mathrm{Exp}_{w_{t-1}^s}^{-1}(w^*)\|_{w_{t-1}^s} = \mathrm{dist}(w_{t-1}^s, w_*) \leq D$ is incorporated. It follows that

$$\begin{aligned}
&\mathbb{E}_{i_t^s}[(\mathrm{dist}(w_t^s, w^*))^2 - (\mathrm{dist}(w_{t-1}^s, w^*))^2] \\
&\leq \mathbb{E}_{i_t^s}\left[\zeta(\mathrm{dist}(w_{t-1}^s, w_t^s))^2 - 2\left\langle -\alpha\xi_t^s, \mathrm{Exp}_{w_{t-1}^s}^{-1}(w^*)\right\rangle_{w_{t-1}^s} + 2c_R D\|\alpha\xi_t^s\|_{w_{t-1}^s}^2\right] \\
&\leq \mathbb{E}_{i_t^s}[(\zeta\nu + 2c_R D)\|\alpha\xi_t^s\|_{w_{t-1}^s}^2] + 2\alpha\left\langle \mathrm{grad} f(w_{t-1}^s), \mathrm{Exp}_{w_{t-1}^s}^{-1}(w^*)\right\rangle_{w_{t-1}^s},
\end{aligned}$$

where the last inequality follows from $\mathbb{E}_{i_t^s}[\xi_t^s] = \mathrm{grad} f(w_{t-1}^s)$. Lemma 5.8, together with the relation $f(w^*) \leq f(w_{t-1}^s)$, yields

$$\begin{aligned}
\left\langle \mathrm{grad} f(w_{t-1}^s), \mathrm{Exp}_{w_{t-1}^s}^{-1}(w^*)\right\rangle_{w_{t-1}^s} &\leq -\frac{\sigma}{2}\left\|\mathrm{Exp}_{w_{t-1}^s}^{-1}(w^*)\right\|_{w_{t-1}^s}^2 \\
&= -\frac{\sigma}{2}(\mathrm{dist}(w_{t-1}^s, w^*))^2.
\end{aligned}$$

We thus obtain, by Lemma 5.13,

$$\begin{aligned}
&\mathbb{E}_{i_t^s}[(\mathrm{dist}(w_t^s, w^*))^2 - (\mathrm{dist}(w_{t-1}^s, w^*))^2] \\
&\leq \mathbb{E}_{i_t^s}[(\zeta\nu + 2c_R D)\|\alpha\xi_t^s\|_{w_{t-1}^s}^2 - \sigma\alpha(\mathrm{dist}(w_{t-1}^s, w^*))^2] \\
&\leq \mathbb{E}_{i_t^s}[4(\zeta\nu + 2c_R D)\alpha^2(\beta^2 + \mu^2C^2\theta^2)(7(\mathrm{dist}(w_{t-1}^s, w^*))^2 + 4(\mathrm{dist}(\tilde{w}^{s-1}, w^*))^2) \\
&\qquad - \sigma\alpha(\mathrm{dist}(w_{t-1}^s, w^*))^2] \\
&= \alpha(28(\zeta\nu + 2c_R D)\alpha(\beta^2 + \mu^2C^2\theta^2) - \sigma)(\mathrm{dist}(w_{t-1}^s, w^*))^2 \\
&\qquad + 16(\zeta\nu + 2c_R D)\alpha^2(\beta^2 + \mu^2C^2\theta^2)(\mathrm{dist}(\tilde{w}^{s-1}, w^*))^2. \qquad (12)
\end{aligned}$$

It follows for the unconditional expectation operator $\mathbb{E}$ that

$$\begin{aligned}
&\mathbb{E}[(\mathrm{dist}(w_t^s, w^*))^2] - \mathbb{E}[(\mathrm{dist}(w_{t-1}^s, w^*))^2] \\
&\leq \alpha(28(\zeta\nu + 2c_R D)\alpha(\beta^2 + \mu^2C^2\theta^2) - \sigma)\mathbb{E}[(\mathrm{dist}(w_{t-1}^s, w^*))^2] \\
&\qquad + 16(\zeta\nu + 2c_R D)\alpha^2(\beta^2 + \mu^2C^2\theta^2)\mathbb{E}[(\mathrm{dist}(\tilde{w}^{s-1}, w^*))^2]. \qquad (13)
\end{aligned}$$

Summing (13) over $t = 1, 2, \ldots, m$ of the inner loop on the sth epoch, we have

$$\begin{aligned}
&\mathbb{E}[(\mathrm{dist}(w_m^s, w^*))^2] - \mathbb{E}[(\mathrm{dist}(w_0^s, w^*))^2] \\
&\leq \alpha(28(\zeta\nu + 2c_R D)\alpha(\beta^2 + \mu^2C^2\theta^2) - \sigma)\sum_{t=1}^{m}\mathbb{E}[(\mathrm{dist}(w_{t-1}^s, w^*))^2] \\
&\qquad + 16m(\zeta\nu + 2c_R D)\alpha^2(\beta^2 + \mu^2C^2\theta^2)\mathbb{E}[(\mathrm{dist}(\tilde{w}^{s-1}, w^*))^2]. \qquad (14)
\end{aligned}$$

Hence, with option II, where we compute $\tilde{w}^s$ as $\tilde{w}^s = w_m^s$, the facts $w_0^s = \tilde{w}^{s-1}$ and $\alpha(28(\zeta\nu + 2c_R D)\alpha(\beta^2 + \mu^2C^2\theta^2) - \sigma) < 0$ imply that

$$\begin{aligned}
&\mathbb{E}[(\mathrm{dist}(\tilde{w}^s, w^*))^2] \\
&\qquad \leq (1 - \sigma\alpha + 4(4m + 7)(\zeta\nu + 2c_R D)(\beta^2 + \mu^2C^2\theta^2)\alpha^2)\mathbb{E}[(\mathrm{dist}(\tilde{w}^{s-1}, w^*))^2].
\end{aligned}$$

On the other hand, if we use option I, where $\tilde{w}^s = g_m(w_1^s, w_2^s, \ldots, w_m^s)$, rearranging (14) yields

$$\begin{aligned}
&\alpha(\sigma - 28(\zeta\nu + 2c_R D)\alpha(\beta^2 + \mu^2 C^2\theta^2)) \sum_{t=1}^{m} \mathbb{E}[(\operatorname{dist}(w_t^s, w^*))^2] \\
&= \alpha(\sigma - 28(\zeta\nu + 2c_R D)\alpha(\beta^2 + \mu^2 C^2\theta^2)) \\
&\quad \times \mathbb{E}\Bigg[\sum_{t=0}^{m-1} (\operatorname{dist}(w_t^s, w^*))^2 + (\operatorname{dist}(w_m^s, w^*))^2 - (\operatorname{dist}(w_0^s, w^*))^2 \Bigg] \\
&\le \mathbb{E}[(\operatorname{dist}(w_0^s, w^*))^2 - (\operatorname{dist}(w_m^s, w^*))^2 \\
&\qquad + 16m(\zeta\nu + 2c_R D)\alpha^2(\beta^2 + \mu^2 C^2\theta^2)(\operatorname{dist}(w_0^s, w^*))^2 \\
&\qquad - \alpha(\sigma - 28(\zeta\nu + 2c_R D)\alpha(\beta^2 + \mu^2 C^2\theta^2))((\operatorname{dist}(w_0^s, w^*))^2 - (\operatorname{dist}(w_m^s, w^*))^2)] \\
&\le (1 - \alpha(\sigma - 28(\zeta\nu + 2c_R D)\alpha(\beta^2 + \mu^2 C^2\theta^2)) \\
&\qquad\qquad + 16m(\zeta\nu + 2c_R D)\alpha^2(\beta^2 + \mu^2 C^2\theta^2))\mathbb{E}[(\operatorname{dist}(w_0^s, w^*))^2] \\
&\le (1 + 16m(\zeta\nu + 2c_R D)\alpha^2(\beta^2 + \mu^2 C^2\theta^2))\mathbb{E}[(\operatorname{dist}(\tilde{w}^{s-1}, w^*))^2].
\end{aligned}$$

Using $\tilde{w}^s = g_m(w_1^s, w_2^s, \ldots, w_m^s)$ and Lemma 5.11, we obtain

$$\mathbb{E}[(\operatorname{dist}(\tilde{w}^s, w^*))^2] \le \frac{4(1 + 16m(\zeta\nu + 2c_R D)(\beta^2 + \mu^2 C^2\theta^2)\alpha^2)}{m\alpha(\sigma - 28(\zeta\nu + 2c_R D)(\beta^2 + \mu^2 C^2\theta^2)\alpha)} \mathbb{E}[(\operatorname{dist}(\tilde{w}^{s-1}, w^*))^2].$$

This completes the proof. ∎

In the above theorem, we note that, from the definitions of β and σ, β can be chosen to be arbitrarily large and σ can be chosen to be arbitrarily small. Therefore, $\alpha = \sigma/56(\zeta\nu + 2c_R D)(\beta^2 + \mu^2 C^2\theta^2)$, e.g., satisfies

$$0 < \alpha(\sigma - 28(\zeta\nu + 2c_R D)(\beta^2 + \mu^2 C^2\theta^2)\alpha) < 1$$

for sufficiently large β and small σ.

In fact, an α satisfying the condition always exists for any values of β and σ. Let

$$\beta' := 28(\zeta\nu + 2c_R D)(\beta^2 + \mu^2 C^2\theta^2).$$

We can analyze the inequality $0 < \alpha(\sigma - 28(\zeta\nu + 2c_R D)(\beta^2 + \mu^2 C^2\theta^2)\alpha) < 1$, which is expressed as $0 < \alpha(\sigma - \beta'\alpha) < 1$, with the condition $\alpha > 0$ written more specifically as

$$\text{(15)} \qquad \begin{cases} 0 < \alpha < \dfrac{\sigma}{\beta'} & \text{if } \sigma^2 - 4\beta' < 0, \\ 0 < \alpha < \dfrac{\sigma}{2\beta'},\ \dfrac{\sigma}{2\beta'} < \alpha < \dfrac{\sigma}{\beta'} & \text{if } \sigma^2 - 4\beta' = 0, \\ 0 < \alpha < \dfrac{\sigma - \sqrt{\sigma^2 - 4\beta'}}{2\beta'},\ \dfrac{\sigma + \sqrt{\sigma^2 - 4\beta'}}{2\beta'} < \alpha < \dfrac{\sigma}{\beta'} & \text{if } \sigma^2 - 4\beta' > 0. \end{cases}$$

Furthermore, we can show that the coefficient δ on the right-hand side of (11), which can be written as $4(7 + 4m\beta'\alpha^2)/7m\alpha(\sigma - \beta'\alpha) =: r_1(\alpha)$ for option I and $1 - \sigma\alpha + (1 + 4m/7)\beta'\alpha^2 =: r_2(\alpha)$ for option II, is less than 1 when m is sufficiently large and α is appropriately chosen. If α is fixed, $r_1(\alpha) \to 16\beta'\alpha/7(\sigma - \beta'\alpha)$ as $m \to \infty$, which is not necessarily less than 1, and $r_2(\alpha) \to \infty$ as $m \to \infty$. Thus,

we again need to specifically analyze an appropriate value of α, which should depend on m.

For option I, by calculating the derivative $r_1'(\alpha)$ of $r_1(\alpha)$ on α, we can show that $r_1(\alpha)$ takes the minimum value at

$$\alpha = \frac{-7\beta' + \sqrt{49\beta'^2 + 28m\beta'\sigma^2}}{4m\beta'\sigma} =: \alpha_*,$$

which satisfies (15) when m is sufficiently large, since $\lim_{m\to\infty} \alpha_* = 0$. Note that we have $r_1'(\alpha_*) = 0$, which yields $4m\beta'\sigma\alpha_*^2 + 14\beta'\alpha_* - 7\sigma = 0$. This relation gives the minimum value $r_1(\alpha_*)$ as

$$r_1(\alpha_*) = \frac{32(\sigma - \beta'\alpha_*)}{2(7\beta' + 2m\sigma^2)\alpha_* - 7\sigma} \to 0 \quad (m \to \infty),$$

where we have used the facts that $\lim_{m\to\infty} \alpha_* = 0$ and $\lim_{m\to\infty} m\alpha_* = \infty$. A simpler choice of $\alpha = 1/\sqrt{m}$ also makes $r_1(\alpha)$ less than 1 if m is sufficiently large since

$$r_1\left(\frac{1}{\sqrt{m}}\right) = \frac{4(7 + 4\beta')}{7\,(\sigma\sqrt{m} - \beta')} \to 0 \quad (m \to \infty).$$

Although $\alpha = 1/\sqrt{m}$ does not achieve the best rate attained by $\alpha = \alpha_*$, this choice is practical because we do not know the exact values of σ, β, or α_* in general.

A similar discussion can be applied to the case of option II. In this case, it is clear that a sufficiently small α satisfies (15) and $r_2(\alpha) < 1$. Furthermore, if $\sigma^2 - 4\beta' \le 0$, $\alpha_* = \sigma/2(1 + 4m/7)\beta'$ satisfies (15) and attains the minimum value of r_2 as

$$r_2(\alpha_*) = 1 - \frac{7\sigma^2}{4(4m + 7)\beta'} < 1.$$

Furthermore, this α_* satisfies (15) if $\sigma^2 - 4\beta' > 0$ and m is sufficiently large.

We have thus shown that a local linear convergence rate is achieved under an appropriate fixed step size if m is sufficiently large, which is the same as standard SVRG in Euclidean space (for nonconvex problems). We can also analyze the rate with decaying step sizes $\alpha_0^s > \alpha_1^s > \cdots > \alpha_m^s$ (at the sth epoch) as

$$\frac{4(1 + 16m(\zeta\nu + 2c_R D)(\beta^2 + \mu^2 C^2\theta^2)(\alpha_0^s)^2)}{m\alpha_m^s(\sigma - 28(\zeta\nu + 2c_R D)(\beta^2 + \mu^2 C^2\theta^2)\alpha_0^s)}$$

for option I. This is larger than

$$\frac{4(1 + 16m(\zeta\nu + 2c_R D)(\beta^2 + \mu^2 C^2\theta^2)(\alpha_0^s)^2)}{m\alpha_0^s(\sigma - 28(\zeta\nu + 2c_R D)(\beta^2 + \mu^2 C^2\theta^2)\alpha_0^s)},$$

which is the coefficient in (11) with the fixed step size $\alpha = \alpha_0^s$. A similar discussion can be applied to the case of option II. Consequently, using decaying step sizes also yields a local convergence, but at a worse rate than with a fixed step size. Both the above guarantees are quite similar to those available for batch gradient algorithms on manifolds. This setup, i.e., hybrid step sizes, follows our two convergence analyses, which first require decaying step sizes to approach a neighborhood of a local minimum and use a fixed step size to achieve faster linear convergence near the solution. As mentioned earlier, we guarantee global convergence and local linear convergence even

if we use decaying step sizes from beginning to end. Therefore, this method is an improved version of decaying step size. Theoretical analysis of the switching between decaying and fixed step sizes is left for future work.

We make one more remark. Although this subsection describes the local convergence analysis with an objective function that is strictly retraction-convex in a sufficiently small neighborhood of a local minimizer, it is also applicable for a global convergence analysis if we assume that the function is globally strictly convex, as in other studies. As we have already discussed, the rate in Theorem 5.14 can lead to discussions on global iteration complexity. The key aspect of our local convergence analysis is that we do not have to assume global convexity to attain the local linear convergence rate.

5.3. Local convergence rate analysis with exponential mapping and parallel translation. In this subsection, we present a local convergence rate analysis of the R-SVRG algorithm with exponential mapping and parallel translation along the geodesics. This is a special case of the result in the previous subsection, where exponential mapping and parallel translation are chosen as retraction and vector transport, respectively. However, in this particular case, we can obtain a stricter rate than in the general case. Since the results are obtained by a similar discussion to that in the previous subsection, we give only a sketch of the proofs.

We obtain the following result as a corollary of the proof of Lemma 5.13, with $R = \mathrm{Exp}$ and $\mathcal{T} = P$.

COROLLARY 5.16. *Suppose the conditions in Lemma* 5.13 *hold and consider Algorithm* 1 *with* $R = \mathrm{Exp}$ *and* $\mathcal{T} = P$*, i.e., the exponential mapping and parallel translation case. Let* β *be the constant in Lemma* 5.13*. Then, the upper bound of* $\mathbb{E}_{i_t^s}[\|\xi_t^s\|^2_{w_{t-1}^s}]$ *is given by*

$$\mathbb{E}_{i_t^s}[\|\xi_t^s\|^2_{w_{t-1}^s}] \le \beta^2(14(\mathrm{dist}(w_{t-1}^s, w^*))^2 + 8(\mathrm{dist}(\tilde{w}^{s-1}, w^*))^2). \tag{16}$$

Proof. Putting $R = \mathrm{Exp}$ and $\mathcal{T} = P$ in the middle of the proof of Lemma 5.13, which is in Appendix A, we obtain

$$\begin{aligned}
&\mathbb{E}_{i_t^s}[\|\xi_t^s\|^2_{w_{t-1}^s}] \\
&\le 2\mathbb{E}_{i_t^s}\left[\left\|\mathrm{grad} f_{i_t^s}(w_{t-1}^s) - P_\gamma^{w_{t-1}^s \leftarrow w^*}(\mathrm{grad} f_{i_t^s}(w^*))\right\|^2_{w_{t-1}^s}\right] \\
&\quad + 2\mathbb{E}_{i_t^s}\left[\left\|P_\gamma^{w_{t-1}^s \leftarrow \tilde{w}^{s-1}}(\mathrm{grad} f_{i_t^s}(\tilde{w}^{s-1})) - P_\gamma^{w_{t-1}^s \leftarrow w^*}(\mathrm{grad} f_{i_t^s}(w^*))\right\|^2_{w_{t-1}^s}\right] \\
&\quad - 2\left\|P_\gamma^{w_{t-1}^s \leftarrow \tilde{w}^{s-1}}(\mathrm{grad} f(\tilde{w}^{s-1}))\right\|^2_{w_{t-1}^s}.
\end{aligned}$$

In a similar manner to Lemma 5.13, we have

$$\begin{aligned}
\mathbb{E}_{i_t^s}[\|\xi_t^s\|^2_{w_{t-1}^s}] \le{}& 2\mathbb{E}_{i_t^s}\left[\left\|\mathrm{grad} f_{i_t^s}(w_{t-1}^s) - P_\gamma^{w_{t-1}^s \leftarrow w^*}(\mathrm{grad} f_{i_t^s}(w^*))\right\|^2_{w_{t-1}^s}\right] \\
&+ 2\mathbb{E}_{i_t^s}\left[\left\|P_\gamma^{w_{t-1}^s \leftarrow \tilde{w}^{s-1}}(\mathrm{grad} f_{i_t^s}(\tilde{w}^{s-1})) - \mathrm{grad} f_{i_t^s}(w_{t-1}^s)\right.\right. \\
&\qquad \left.\left. + \mathrm{grad} f_{i_t^s}(w_{t-1}^s) - P_\gamma^{w_{t-1}^s \leftarrow w^*}(\mathrm{grad} f_{i_t^s}(w^*))\right\|^2_{w_{t-1}^s}\right]
\end{aligned}$$

$$
\begin{aligned}
&\le 2\mathbb{E}_{i_t^s}\left[\left\|\operatorname{grad} f_{i_t^s}(w_{t-1}^s) - P_\gamma^{w_{t-1}^s \leftarrow w^*}(\operatorname{grad} f_{i_t^s}(w^*))\right\|_{w_{t-1}^s}^2\right] \\
&\quad + 4\mathbb{E}_{i_t^s}\left[\left\|P_\gamma^{w_{t-1}^s \leftarrow \tilde{w}^{s-1}}(\operatorname{grad} f_{i_t^s}(\tilde{w}^{s-1})) - \operatorname{grad} f_{i_t^s}(w_{t-1}^s)\right\|_{w_{t-1}^s}^2\right] \\
&\quad + 4\mathbb{E}_{i_t^s}\left[\left\|\operatorname{grad} f_{i_t^s}(w_{t-1}^s) - P_\gamma^{w_{t-1}^s \leftarrow w^*}(\operatorname{grad} f_{i_t^s}(w^*))\right\|_{w_{t-1}^s}^2\right] \\
&\le \beta^2(6(\operatorname{dist}(w_{t-1}^s, w^*))^2 + 4(\operatorname{dist}(\tilde{w}^{s-1}, w_{t-1}^s))^2) \\
&\le \beta^2(6(\operatorname{dist}(w_{t-1}^s, w^*))^2 + 4(\operatorname{dist}(\tilde{w}^{s-1}, w^*) + \operatorname{dist}(w^*, w_{t-1}^s))^2) \\
&\le \beta^2(6(\operatorname{dist}(w_{t-1}^s, w^*))^2 + 8(\operatorname{dist}(\tilde{w}^{s-1}, w^*))^2 + 8(\operatorname{dist}(w^*, w_{t-1}^s))^2) \\
&= \beta^2(14(\operatorname{dist}(w_{t-1}^s, w^*))^2 + 8(\operatorname{dist}(\tilde{w}^{s-1}, w^*))^2).
\end{aligned}
$$

This completes the proof. ∎

COROLLARY 5.17. *Suppose the conditions in Theorem* 5.14 *hold, except that a positive number* α *satisfies* $0 < \alpha(\sigma - 14\zeta\beta^2\alpha) < 1$*, and consider Algorithm* 1 *with a fixed step size* $\alpha_t^s := \alpha$ *for the exponential mapping and parallel translation case. For any sequence* $\{\tilde{w}^s\}$ *generated by the algorithm, there exists* $K > 0$ *such that, for all* $s > K$*,*

$$
\mathbb{E}[(\operatorname{dist}(\tilde{w}^s, w^*))^2] \le \delta_0 \mathbb{E}[(\operatorname{dist}(\tilde{w}^{s-1}, w^*))^2],
$$

where

$$
\delta_0 := \begin{cases} \dfrac{4\left(1 + 8m\zeta\beta^2\alpha^2\right)}{m\alpha\left(\sigma - 14\zeta\beta^2\alpha\right)} & \text{with option I,} \\ 1 - \sigma\alpha + (8m + 14)\zeta\beta^2\alpha^2 & \text{with option II.} \end{cases}
$$

Proof. Note that the constants in Theorem 5.14 are $c_R = \theta = 0$ and $\mu = \nu = 1$ in this case. By using (16) instead of (10), we obtain

$$
\begin{aligned}
&\mathbb{E}_{i_t^s}[(\operatorname{dist}(w_t^s, w^*))^2 - (\operatorname{dist}(w_{t-1}^s, w^*))^2] \\
&\qquad \le \alpha(14\zeta\alpha\beta^2 - \sigma)(\operatorname{dist}(w_{t-1}^s, w^*))^2 + 8\zeta\alpha^2\beta^2(\operatorname{dist}(\tilde{w}^{s-1}, w^*))^2
\end{aligned}
$$

instead of (12). Summing over $t = 1, 2, \ldots, m$ of the inner loop on the sth epoch, we have

$$
\begin{aligned}
&\mathbb{E}[(\operatorname{dist}(w_m^s, w^*))^2] - \mathbb{E}[(\operatorname{dist}(w_0^s, w^*))^2] \\
&\qquad \le \alpha(14\zeta\alpha\beta^2 - \sigma)\sum_{t=1}^{m} \mathbb{E}[(\operatorname{dist}(w_{t-1}^s, w^*))^2] + 8m\zeta\alpha^2\beta^2\mathbb{E}[(\operatorname{dist}(\tilde{w}^{s-1}, w^*))^2].
\end{aligned}
$$

A similar discussion as in the proof of Theorem 5.14 yields the claimed convergence rates. ∎

6. Numerical comparisons. This section compares the performance of R-SVRG(+) (with option II) with that of the Riemannian extension of SGD, i.e., R-SGD, where the Riemannian stochastic gradient algorithm uses $\operatorname{grad} f_{i_t^s}(w_{t-1}^s)$ instead of ξ_t^s in (3). We also make a comparison with R-SD, which is the Riemannian steepest descent algorithm with backtracking line search [1, Chapter 4]. We consider both *fixed* step size and *decaying* step size sequences. The decaying step size sequence

uses the decay $\alpha_k = \alpha_0(1 + \alpha_0\lambda\lfloor k/m_s \rfloor)^{-1}$, where k is the number of iterations. We select some values of α_0 and consider three values of λ (10^{-1}, 10^{-2}, and 10^{-3}). In addition, since the global convergence analysis needs a decaying step size condition and the local convergence rate analysis holds for a fixed step size (sections 4 and 5), we consider a *hybrid* step size sequence that follows the decaying step size before the s_{TH} epoch and subsequently switches to a fixed step size. All the experiments use $m_s = 5N$ by following [15]. In all the figures, the x-axis represents the computational cost measured by the number of gradient computations divided by N. The algorithms are initialized randomly and are terminated when either the stochastic gradient norm is below 10^{-8} or the number of iterations exceeds a predefined threshold. It should be noted that all the results except those of R-SD are the best-tuned results. All the simulations were performed in MATLAB on a 2.6 GHz Intel Core i7 machine with 16 GB of RAM. Hereinafter, this paper addresses three problems on the SPD manifold and the Grassmann manifold. In all the problems, full gradient methods, e.g., the steepest descent algorithm, become prohibitively computationally expensive when N is extremely large. The stochastic gradient approach is a promising way to achieve scalability.

6.1. Problem on the SPD manifold and simulation results. We first consider the Riemannian centroid problem on the SPD manifold.

SPD manifold $\mathcal{S}_{++}^d$ and optimization tools. We designate the space of $d \times d$ SPD matrices as the SPD manifold, $\mathcal{S}_{++}^d$. If we endow $\mathcal{S}_{++}^d$ with the affine-invariant Riemannian metric (AIRM) [25] defined by $\langle \xi_{\mathbf{X}}, \eta_{\mathbf{X}} \rangle_{\mathbf{X}} = \mathrm{trace}(\xi_{\mathbf{X}}\mathbf{X}^{-1}\eta_{\mathbf{X}}\mathbf{X}^{-1})$ for $\xi_{\mathbf{X}}, \eta_{\mathbf{X}} \in T_{\mathbf{X}}\mathcal{S}_{++}^d$ at $\mathbf{X} \in \mathcal{S}_{++}^d$, then the SPD manifold $\mathcal{S}_{++}^d$ forms a Riemannian manifold. The exponential mapping is written as

$$\mathrm{Exp}_{\mathbf{X}}(\xi_{\mathbf{X}}) = \mathbf{X}^{1/2} \exp(\mathbf{X}^{-1/2}\xi_{\mathbf{X}}\mathbf{X}^{-1/2})\mathbf{X}^{1/2}$$

for any $\xi_{\mathbf{X}}$ and $\mathbf{X}$. The parallel translation of $\xi_{\mathbf{X}}$ along $\eta_{\mathbf{X}}$ on $\mathcal{S}_{++}^d$ is given by $P_{\eta_{\mathbf{X}}}(\xi_{\mathbf{X}}) = \mathbf{X}^{1/2}\mathbf{Y}\mathbf{X}^{-1/2}\xi_{\mathbf{X}}\mathbf{X}^{-1/2}\mathbf{Y}\mathbf{X}^{1/2}$, where $\mathbf{Y} = \exp(\mathbf{X}^{-1/2}\eta_{\mathbf{X}}\mathbf{X}^{-1/2}/2)$. The logarithm map of $\mathbf{Y}$ at $\mathbf{X}$ is described as

$$\mathrm{Log}_{\mathbf{X}}(\mathbf{Y}) = \mathbf{X}^{1/2} \log(\mathbf{X}^{-1/2}\mathbf{Y}\mathbf{X}^{-1/2})\mathbf{X}^{1/2} = \log(\mathbf{Y}\mathbf{X}^{-1})\mathbf{X}.$$

The exponential mapping and the parallel translation above are computationally expensive. Therefore, the following efficient retraction is proposed [14]:

$$R_{\mathbf{X}}(\xi_{\mathbf{X}}) = \mathbf{X} + \xi_{\mathbf{X}} + \frac{1}{2}\xi_{\mathbf{X}}\mathbf{X}^{-1}\xi_{\mathbf{X}}. \tag{17}$$

This maps $\xi_{\mathbf{X}}$ onto $\mathcal{S}_{++}^d$ for all $\xi_{\mathbf{X}} \in T_{\mathbf{X}}\mathcal{S}_{++}^d$. Huang et al. proposed an efficient isometric vector transport [11, 33] defined as

$$\mathcal{T}_{S_\eta}\xi_{\mathbf{X}} = B_{\mathbf{Y}}B_{\mathbf{X}}^{\flat}\xi_{\mathbf{X}}, \tag{18}$$

where $\mathbf{Y} = R_{\mathbf{X}}(\xi_{\mathbf{X}})$ and $a^\flat$ denotes the flat of $a \in T_w\mathcal{M}$, i.e., $a^\flat \colon T_w\mathcal{M} \to \mathbb{R} \colon v \mapsto \langle a, v \rangle_w$. $B_{\mathbf{X}}$ and $B_{\mathbf{Y}}$ are the orthonormal bases of $T_{\mathbf{X}}\mathcal{S}_{++}^d$ and $T_{\mathbf{Y}}\mathcal{S}_{++}^d$, respectively, where the basis is calculated based on the Cholesky decomposition. Consequently, the implementation of our algorithm for this particular problem uses the retraction (17) and vector transport (18), which satisfy the requirements in the convergence analyses in sections 4 and 5.

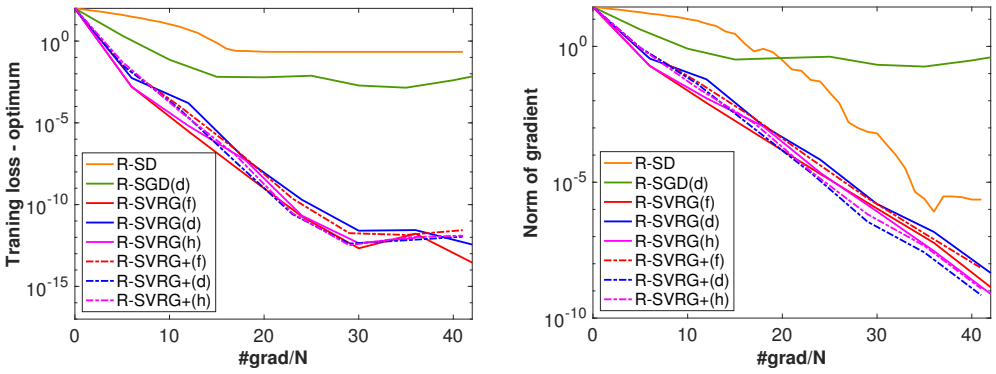
Riemannian centroid problem. We first evaluate the proposed algorithm in the Riemannian centroid problem on $\mathcal{S}_{++}^d$, which is frequently used for computer vision problems such as visual object categorization and pose categorization [13]. Given N points on $\mathcal{S}_{++}^d$ with matrix representations $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_N$, the Riemannian centroid is derived from the solution to the problem

$$\min_{\mathbf{C} \in \mathcal{S}_{++}^d} \frac{1}{2N} \sum_{n=1}^{N} (\mathrm{dist}(\mathbf{C}, \mathbf{X}_n))^2,$$

where $\mathrm{dist}(\mathbf{A}, \mathbf{B}) = \|\log(\mathbf{A}^{-1/2}\mathbf{B}\mathbf{A}^{-1/2})\|_F$ represents the distance along the corresponding geodesic between the two points $\mathbf{A}, \mathbf{B} \in \mathcal{S}_{++}^d$ with respect to the AIRM. The gradient of the loss function is computed as

$$\frac{1}{N} \sum_{n=1}^{N} -\log(\mathbf{X}_n \mathbf{C}^{-1})\mathbf{C}.$$

Parts (a) and (b) of Figure 1 show the results of the optimality gap and the norm of the gradient, respectively, where $N = 1000$ and $d = 3$. The choices of α_0 are $\{10^{-3}, 2\times10^{-3}, 4\times10^{-3}, 6\times10^{-3}, 8\times10^{-3}, 10^{-2}, 2\times10^{-2}, \ldots, 10^{-1}\}$.[1] s_{TH} and the batch size are fixed to 3 and 1, respectively. The maximum number of iterations is 10 for R-SVRG(+) and 60 for the others. The optimality gap evaluates the performance against the minimum loss, which is obtained by R-SD with high precision in advance. We can see from the figures that R-SVRG(+) outperforms R-SGD in terms of the gradient counts and exhibits much faster convergence than R-SD, as expected.



(a) Optimality gap (b) Norm of gradient

FIG. 1. *Performance evaluations on the Riemannian centroid problem. In the legends of the figures, (f), (d), and (h) denote fixed, decaying, and hybrid step sizes, respectively. The parameters are chosen as follows. R-SGD(d):* $\alpha_0 = 0.002$, $\lambda = 0.1$; *R-SVRG(f):* $\alpha_0 = 0.008$; *R-SVRG(d):* $\alpha_0 = 0.1$, $\lambda = 0.1$; *R-SVRG(h):* $\alpha_0 = 0.01$, $\lambda = 0.001$; *R-SVRG+(f):* $\alpha_0 = 0.01$; *R-SVRG+(d):* $\alpha_0 = 0.006$, $\lambda = 0.01$; *R-SVRG+(h):* $\alpha_0 = 0.004$, $\lambda = 0.001$.

6.2. Problems on the Grassmann manifold and simulation results. We focus on two popular problems on the Grassmann manifold: PCA and low-rank matrix completion problems.

[1] Color figures are available in the online version of this paper.

Grassmann manifold and optimization tools. An element on the Grassmann manifold is represented by a $d \times r$ orthogonal matrix $\mathbf{U}$ with orthonormal columns, i.e., $\mathbf{U}^T\mathbf{U} = \mathbf{I}$. Two orthogonal matrices represent the same element on the Grassmann manifold if they are related by right multiplication of an $r \times r$ orthogonal matrix $\mathbf{O} \in \mathcal{O}(r)$, where $\mathcal{O}(r)$ is the orthogonal group of order r. Equivalently, an element of the Grassmann manifold is identified with a set of $d \times r$ orthogonal matrices $[\mathbf{U}] := \{\mathbf{UO} \mid \mathbf{O} \in \mathcal{O}(r)\}$. Thus, $\mathrm{Gr}(r,d) := \mathrm{St}(r,d)/\mathcal{O}(r)$, where $\mathrm{St}(r,d)$ is the Stiefel manifold, which is the set of matrices of size $d \times r$ with orthonormal columns. The Grassmann manifold has the structure of a Riemannian quotient manifold [1, section 3.4]. The exponential mapping for the Grassmann manifold from $\mathbf{U}(0) := \mathbf{U} \in \mathrm{Gr}(r,d)$ in the direction of $\xi \in T_{\mathbf{U}(0)}\mathrm{Gr}(r,d)$ is given in a closed form as [1, section 5.4]

$$\mathbf{U}(t) = [\mathbf{U}(0)\mathbf{V} \quad \mathbf{W}] \begin{bmatrix} \cos t\Sigma \\ \sin t\Sigma \end{bmatrix} \mathbf{V}^T,$$

where $\xi = \mathbf{W}\Sigma\mathbf{V}^T$ is the rank-r singular value decomposition of ξ. The $\cos(\cdot)$ and $\sin(\cdot)$ operations are only on the diagonal entries. The parallel translation of $\zeta \in T_{\mathbf{U}(0)}\mathrm{Gr}(r,d)$ on the Grassmann manifold along $\gamma(t)$ with $\dot{\gamma}(0) = \mathbf{W}\Sigma\mathbf{V}^T$ is given in a closed form by

$$\zeta(t) = \left([\mathbf{U}(0)\mathbf{V} \quad \mathbf{W}] \begin{bmatrix} -\sin t\Sigma \\ \cos t\Sigma \end{bmatrix} \mathbf{W}^T + (\mathbf{I} - \mathbf{W}\mathbf{W}^T) \right) \zeta.$$

The logarithm map of $\mathbf{U}(t)$ at $\mathbf{U}(0)$ on the Grassmann manifold is given by

$$\xi = \log_{\mathbf{U}(0)}(\mathbf{U}(t)) = \mathbf{W} \arctan(\Sigma)\mathbf{V}^T,$$

where the rank-r singular value decomposition of

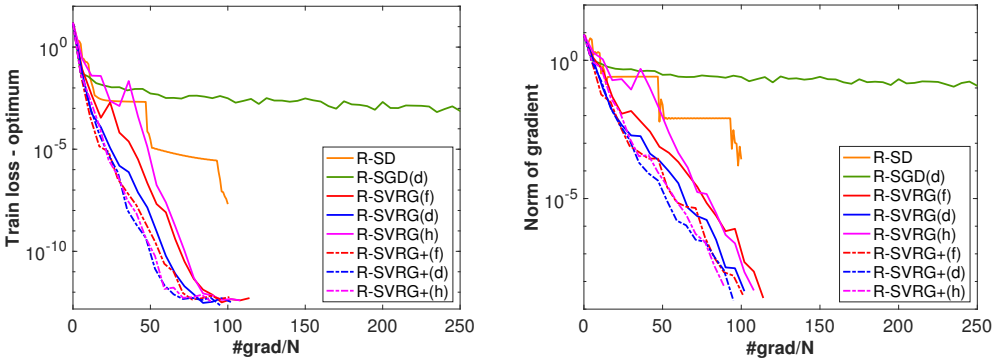
$$(\mathbf{U}(t) - \mathbf{U}(0)\mathbf{U}(0)^T\mathbf{U}(t))(\mathbf{U}(0)^T\mathbf{U}(t))^{-1}$$

is $\mathbf{W}\Sigma\mathbf{V}^T$. It should be noted that this experiment evaluates the projection-based vector transport and the QR-decomposition-based retraction, which do not satisfy the conditions in sections 4 and 5 but are computationally efficient. The intention here is to show that our algorithm performs well empirically without using the specific vector transport.

The PCA problem. Given an orthonormal matrix projector $\mathbf{U} \in \mathrm{St}(r,d)$, the PCA problem is to minimize the sum of the squared residual errors between the projected data points and the original data as

$$\min_{\mathbf{U} \in \mathrm{St}(r,d)} \quad \frac{1}{N} \sum_{n=1}^{N} \|\boldsymbol{x}_n - \mathbf{U}\mathbf{U}^T\boldsymbol{x}_n\|_2^2, \tag{19}$$

where $\boldsymbol{x}_n$ is a data vector of size $d \times 1$. Problem (19) is equivalent to maximizing the function $\frac{1}{N}\sum_{n=1}^{N} \boldsymbol{x}_n^T\mathbf{U}\mathbf{U}^T\boldsymbol{x}_n$. Here, the critical points in the space $\mathrm{St}(r,d)$ are not isolated because the cost function remains unchanged under the group action $\mathbf{U} \mapsto \mathbf{UO}$ for any orthogonal matrices $\mathbf{O}$ of size $r \times r$. Consequently, problem (19) is reformulated as an optimization problem on the Grassmann manifold $\mathrm{Gr}(r,d)$.



(a) Optimality gap (b) Norm of gradient

FIG. 2. *Performance evaluations on the PCA problem. In the legends of the figures, (f), (d), and (h) denote fixed, decaying, and hybrid step sizes, respectively. The parameters are chosen as follows. R-SGD(d):* $\alpha_0 = 0.009$, $\lambda = 0.1$; *R-SVRG(f):* $\alpha_0 = 0.001$; *R-SVRG(d):* $\alpha_0 = 0.001$, $\lambda = 0.001$; *R-SVRG(h):* $\alpha_0 = 0.004$, $\lambda = 0.01$; *R-SVRG+(f):* $\alpha_0 = 0.001$; *R-SVRG+(d):* $\alpha_0 = 0.002$, $\lambda = 0.01$; *R-SVRG+(h):* $\alpha_0 = 0.002$, $\lambda = 0.01$.

Figures 2(a) and (b) show the optimality gap and gradient norm, respectively, where $N = 10000$, $d = 20$, and $r = 5$. The choices of α_0 are $\{10^{-3}, 2\times10^{-3}, \dots, 10^{-2}\}$. The minimum loss for the optimality gap evaluation is obtained by the MATLAB function `pca`. s_{TH} and the batch size are fixed to 5 and 10, respectively. The maximum number of iterations is 16 for R-SVRG(+) and 100 for the others. From Figure 2(a), we can observe that, of R-SVRG and R-SVRG+, the latter shows superior performance for all the step size sequences. In Figure 2(b), the gradient norm of SGD remains at higher values, while those of R-SVRG and R-SVRG+ converge to lower values in all the cases.

Low-rank matrix completion. The matrix completion problem amounts to completing an incomplete matrix $\mathbf{X}$, say of size $d \times N$, from a small number of entries by assuming a low-rank model for the matrix. If Ω is the set of indices for which we know the entries in $\mathbf{X}$, the rank-r matrix completion problem amounts to solving the problem

$$\min_{\mathbf{U}\in\mathbb{R}^{d\times r},\ \mathbf{A}\in\mathbb{R}^{r\times N}} \|\mathcal{P}_\Omega(\mathbf{UA}) - \mathcal{P}_\Omega(\mathbf{X})\|_F^2, \tag{20}$$

where the operator $\mathcal{P}_\Omega$ acts as $\mathcal{P}_\Omega(\mathbf{X}_{ij}) = \mathbf{X}_{ij}$ if $(i,j) \in \Omega$ and $\mathcal{P}_\Omega(\mathbf{X}_{ij}) = 0$ otherwise. This is called the orthogonal sampling operator and is a mathematically convenient way to represent the subset of known entries. Partitioning $\mathbf{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_n]$, problem (20) is equivalent to

$$\min_{\mathbf{U}\in\mathbb{R}^{d\times r},\ \boldsymbol{a}_n\in\mathbb{R}^r} \frac{1}{N}\sum_{n=1}^{N} \|\mathcal{P}_{\Omega_n}(\mathbf{U}\boldsymbol{a}_n) - \mathcal{P}_{\Omega_n}(\boldsymbol{x}_n)\|_2^2, \tag{21}$$

where $\boldsymbol{x}_n \in \mathbb{R}^d$ and the operator $\mathcal{P}_{\Omega_n}$ is the sampling operator for the nth column. Given $\mathbf{U}$, $\boldsymbol{a}_n$ in (21) admits a closed-form solution. Consequently, problem (21) depends only on the column space of $\mathbf{U}$ and is on the Grassmann manifold [4].

The proposed algorithms are also compared with Grouse [4], a state-of-the-art stochastic descent algorithm on the Grassmann manifold. We first consider a synthetic dataset with $N = 5000$ and $d = 500$ with rank $r = 5$. The algorithms are initialized randomly as suggested in [20]. The 10 choices of α_0 are $\{10^{-3}, 2\times 10^{-3}, \ldots, 10^{-2}\}$ for R-SGD and R-SVRG(+), and $\{0.1, 0.2, \ldots, 1.0\}$ for Grouse. This instance considers the loss on a test set Γ, which differs from training set Ω. We also impose an exponential decay of the singular values. The condition number (CN) of a matrix is the ratio of the largest to the smallest singular values of the matrix. The oversampling ratio (OS) determines the number of entries that are known. This instance uses CN $= 5$ and OS $= 5$. An OS of 5 implies that $5(N + d - r)r$ randomly and uniformly selected entries are known a priori among the total Nd entries. Figure 3(a) shows the results of loss on the test set Γ. These results show the superior performance of our proposed algorithms.

Next, we consider Jester dataset 1 [10], consisting of ratings of 100 jokes by 24983 users. Each rating is a real number from -10 to 10. We randomly extract two ratings per user as the training set Ω and test set Γ. The algorithms are run by fixing the rank to $r = 5$ with random initialization. α_0 is chosen from $\{10^{-6}, 2\times 10^{-6}, \ldots, 10^{-5}\}$ for SGD and SVRG(+) and $\{10^{-3}, 2\times 10^{-3}, \ldots, 10^{-2}\}$ for Grouse. Figure 3(b) shows the superior performance of R-SVRG(+) on the test set of the Jester dataset.

As a final test, we compare the algorithms on the MovieLens-1M dataset available at https://grouplens.org/datasets/movielens/1m/, which has a million ratings corresponding to 6040 users and 3952 movies. α_0 is chosen from $\{10^{-5}, 2\times 10^{-5}, \ldots, 10^{-4}\}$. Figure 3(c) shows the results on the test set for all the algorithms except Grouse, which faces issues with convergence on this data set. R-SVRG(+) shows much faster convergence than the others, and R-SVRG is better than R-SVRG+ in terms of the final test loss for all step size algorithms.

7. Conclusion. We proposed a Riemannian stochastic variance reduced gradient (R-SVRG) algorithm with retraction and vector transport, which includes the algorithm with exponential mapping and parallel translation as a special case. The proposed algorithm stems from the variance reduced gradient algorithm in Euclidean space, but here it has been extended to Riemannian manifolds. The main challenges of averaging, adding, and subtracting multiple gradients on a Riemannian manifold were handled with a vector transport. We proved that R-SVRG generates globally convergent sequences with a decaying step size and is locally linearly convergent with a fixed step size under some natural assumptions. Numerical comparisons of problems on the SPD manifold and the Grassmann manifold indicated the superior performance of R-SVRG on various benchmarks.

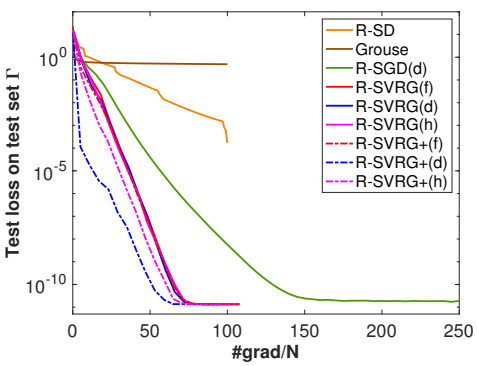
Appendix A. Proofs of the lemmas. In this section, we present complete proofs of Lemmas 5.11–5.13.

Proof of Lemma 5.11. From the triangle inequality and $(a+b)^2 \le 2a^2 + 2b^2$ for real numbers a and b, we have, for $i = 1, 2, \ldots, m$,

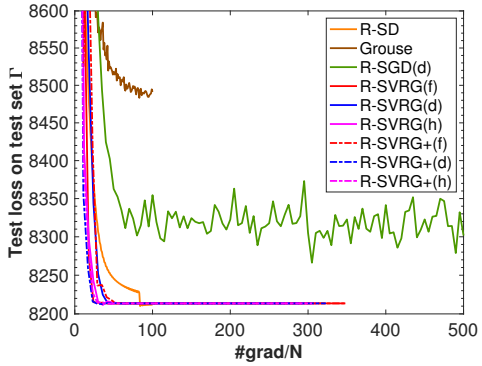
$$(\mathrm{dist}(p, w))^2 \le (\mathrm{dist}(p, w_i) + \mathrm{dist}\,(w_i, w))^2 \le 2(\mathrm{dist}(p, w_i))^2 + 2\,(\mathrm{dist}\,(w_i, w))^2\,.$$

Since w is the Riemannian centroid of $w_1, w_2, \ldots, w_m$, it holds that

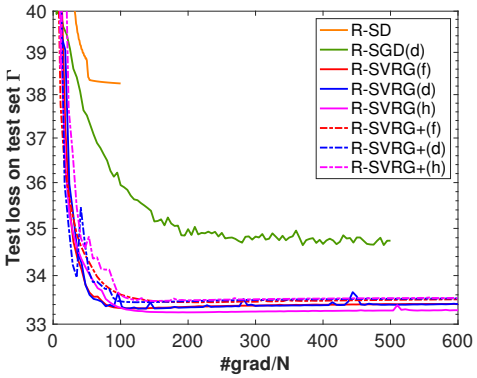
$$\sum_{i=1}^{m} (\mathrm{dist}\,(w, w_i))^2 \le \sum_{i=1}^{m} (\mathrm{dist}(p, w_i))^2.$$



(a) Test loss (synthetic)



(b) Test loss (Jester)



(c) Test loss (MovieLens-1M)

FIG. 3. *Performance evaluations on the low-rank matrix completion problem. In the legends of the figures, (f), (d), and (h) denote fixed, decaying, and hybrid step sizes, respectively. The parameters are chosen as follows. Grouse:* (a) $\alpha_0 = 1$; (b) $\alpha_0 = 0.001$; *R-SGD(d):* (a) $\alpha_0 = 0.001$, $\lambda = 0.01$; (b) $\alpha_0 = 10^{-6}$, $\lambda = 0.1$; (c) $\alpha_0 = 10^{-5}$, $\lambda = 0.001$; *R-SVRG(f):* (a) $\alpha_0 = 0.002$; (b) $\alpha_0 = 5 \times 10^{-6}$; (c) $\alpha_0 = 5 \times 10^{-5}$; *R-SVRG(d):* (a) $\alpha_0 = 0.004$, $\lambda = 0.01$; (b) $\alpha_0 = 7 \times 10^{-6}$, $\lambda = 0.001$; (c) $\alpha_0 = 4 \times 10^{-5}$, $\lambda = 0.001$; *R-SVRG(h):* (a) $\alpha_0 = 0.003$, $\lambda = 0.01$; (b) $\alpha_0 = 6 \times 10^{-6}$, $\lambda = 0.01$; (c) $\alpha_0 = 4 \times 10^{-5}$, $\lambda = 0.01$; *R-SVRG+(f):* (a) $\alpha_0 = 0.002$; (b) $\alpha_0 = 6 \times 10^{-6}$; (c) $\alpha_0 = 3 \times 10^{-5}$; *R-SVRG+(d):* (a) $\alpha_0 = 0.01$, $\lambda = 0.01$; (b) $\alpha_0 = 6 \times 10^{-6}$, $\lambda = 0.01$; (c) $\alpha_0 = 5 \times 10^{-5}$, $\lambda = 0.001$; *R-SVRG+(h):* (a) $\alpha_0 = 0.003$, $\lambda = 0.01$; (b) $\alpha_0 = 7 \times 10^{-6}$, $\lambda = 0.001$; (c) $\alpha_0 = 5 \times 10^{-5}$, $\lambda = 0.1$.

It then follows that

$$m\left(\operatorname{dist}(p,w)\right)^2 \leq 2\sum_{i=1}^{m}(\operatorname{dist}(p,w_i))^2 + 2\sum_{i=1}^{m}\left(\operatorname{dist}\left(w_i,w\right)\right)^2 \leq 4\sum_{i=1}^{m}(\operatorname{dist}(p,w_i))^2.$$

This completes the proof. ∎

Proof of Lemma 5.12. We first show that $\|\operatorname{grad} f_n\|$ for any $n \in \{1,2,\ldots,N\}$ is upper bounded in Ω when Ω is sufficiently small. Since Ω is sufficiently small, it is contained in a set $U \subset \mathcal{M}$ diffeomorphic to an open set $U' \subset \mathbb{R}^{\dim \mathcal{M}}$ by a chart $\phi\colon U \to U'$. Consider a sufficiently small closed ball B in U' centered at $\phi(w^*)$ such that $\phi^{-1}(B) \subset U$. Then, w^* is in $\phi^{-1}(B)$. Note that B is compact and ϕ is a diffeomorphism. Hence, $\phi^{-1}(B)$ is also compact. Replacing Ω with this $\phi^{-1}(B)$ if

necessary, we can assume that Ω is compact. Therefore, $\|\mathrm{grad} f_n\|$ is upper bounded in Ω. In this proof, let C denote a constant such that $\|\mathrm{grad} f_n(z)\|_z \leq C$ for all $n \in \{1, 2, \ldots, N\}$ and $z \in \Omega$.

Let $\tilde{\eta}_{t-1}^s$ satisfy $R_{\tilde{w}^{s-1}}(\tilde{\eta}_{t-1}^s) = w_{t-1}^s$. The definition of ξ_t^s, (3), and the triangle inequality yield

$$\begin{aligned}
\|\xi_t^s\|_{w_{t-1}^s} &= \|\mathrm{grad} f_{i_t^s}(w_{t-1}^s) - \mathcal{T}_{\tilde{\eta}_{t-1}^s}(\mathrm{grad} f_{i_t^s}(\tilde{w}^{s-1}) - \mathrm{grad} f(\tilde{w}^{s-1}))\|_{w_{t-1}^s} \\
&\leq \|\mathrm{grad} f_{i_t^s}(w_{t-1}^s) - \mathcal{T}_{\tilde{\eta}_{t-1}^s}(\mathrm{grad} f_{i_t^s}(\tilde{w}^{s-1}))\|_{w_{t-1}^s} + \|\mathcal{T}_{\tilde{\eta}_{t-1}^s}(\mathrm{grad} f(\tilde{w}^{s-1}))\|_{w_{t-1}^s} \\
&= \|\mathrm{grad} f_{i_t^s}(w_{t-1}^s) - P^{w_{t-1}^s \leftarrow \tilde{w}^{s-1}}(\mathrm{grad} f_{i_t^s}(\tilde{w}^{s-1}))\|_{w_{t-1}^s} \\
&\quad + \|P^{w_{t-1}^s \leftarrow \tilde{w}^{s-1}}(\mathrm{grad} f_{i_t^s}(\tilde{w}^{s-1})) - \mathcal{T}_{\tilde{\eta}_{t-1}^s}(\mathrm{grad} f_{i_t^s}(\tilde{w}^{s-1}))\|_{w_{t-1}^s} \\
&\quad + \|\mathrm{grad} f(\tilde{w}^{s-1})\|_{\tilde{w}^{s-1}},
\end{aligned}$$

where $\gamma_{t-1}^s(\tau) = R_{\tilde{w}^{s-1}}(\tau \tilde{\eta}_{t-1}^s)$ and P is the parallel translation operator along curve γ_{t-1}^s. The three terms on the right-hand side above are bounded as follows. For the first term, it follows from Lemma 5.6 that

$$\|\mathrm{grad} f_{i_t^s}(w_{t-1}^s) - P^{w_{t-1}^s \leftarrow \tilde{w}^{s-1}}(\mathrm{grad} f_{i_t^s}(\tilde{w}^{s-1}))\|_{w_{t-1}^s} \leq \beta \,\mathrm{dist}(w_{t-1}^s, \tilde{w}^{s-1}) \leq 2\beta r,$$

where β is as in the lemma. The second term can be evaluated from Lemmas 5.7 and 5.9 as

$$\begin{aligned}
&\|P^{w_{t-1}^s \leftarrow \tilde{w}^{s-1}}(\mathrm{grad} f_{i_t^s}(\tilde{w}^{s-1})) - \mathcal{T}_{\tilde{\eta}_{t-1}^s}(\mathrm{grad} f_{i_t^s}(\tilde{w}^{s-1}))\|_{w_{t-1}^s} \\
&\qquad \leq \theta \|\mathrm{grad} f_{i_t^s}(\tilde{w}^{s-1})\|_{\tilde{w}^{s-1}} \|\tilde{\eta}_{t-1}^s\|_{\tilde{w}^{s-1}} \leq \theta C \mu \,\mathrm{dist}(w_{t-1}^s, \tilde{w}^{s-1}) \leq 2\theta C \mu r,
\end{aligned}$$

where θ and μ are as in the lemmas. From Lemma 5.6, we have

$$\begin{aligned}
\|\mathrm{grad} f(\tilde{w}^{s-1})\|_{\tilde{w}^{s-1}} &= \|\mathrm{grad} f(\tilde{w}^{s-1}) - P_*^{\tilde{w}^{s-1} \leftarrow w^*}(\mathrm{grad} f(w^*))\|_{\tilde{w}^{s-1}} \\
&\leq \beta \,\mathrm{dist}(\tilde{w}^{s-1}, w^*) \leq \beta r
\end{aligned}$$

for the third term, where P_* is the parallel translation along curve $\tilde{\gamma}_*^{s-1}$ defined by $\tilde{\gamma}_*^{s-1}(\tau) = R_{w^*}(\tau \tilde{\eta}_*^{s-1})$ with $\tilde{\eta}_*^{s-1}$ satisfying $R_{w^*}(\tilde{\eta}_*^{s-1}) = \tilde{w}^{s-1}$. Therefore, we have

$$\|\xi_t^s\|_{w_{t-1}^s} \leq r(3\beta + 2\theta C \mu) < \varepsilon$$

if we choose a sufficiently small r such that $r < \varepsilon/(3\beta + 2\theta C\mu)$. ∎

Proof of Lemma 5.13. Let $\eta_{t-1}^{*s} \in T_{w^*}\mathcal{M}$ and $\tilde{\eta}_{t-1}^s \in T_{\tilde{w}^{s-1}}\mathcal{M}$ satisfy

$$R_{w^*}(\eta_{t-1}^{*s}) = w_{t-1}^s \quad \text{and} \quad R_{\tilde{w}^{s-1}}(\tilde{\eta}_{t-1}^s) = w_{t-1}^s,$$

respectively. Let $P^{w \leftarrow z}$ be a parallel translation along the curve $R_z(\tau\eta)$, where $R_z(\eta) = w$. By Lemmas 5.6 and 5.7, the upper bound of $\mathbb{E}_{i_t^s}[\|\xi_t^s\|_{w_{t-1}^s}^2]$ in terms of the distance of w_{t-1}^s and $\tilde{w}^{s-1}$ from w^* is computed as

$$\begin{aligned}
&\mathbb{E}_{i_t^s}[\|\xi_t^s\|_{w_{t-1}^s}^2] \\
&= \mathbb{E}_{i_t^s}[\|(\mathrm{grad} f_{i_t^s}(w_{t-1}^s) - \mathcal{T}_{\eta_{t-1}^{*s}}(\mathrm{grad} f_{i_t^s}(w^*))) \\
&\qquad + (\mathcal{T}_{\eta_{t-1}^{*s}}(\mathrm{grad} f_{i_t^s}(w^*)) - \mathcal{T}_{\tilde{\eta}_{t-1}^s}(\mathrm{grad} f_{i_t^s}(\tilde{w}^{s-1})) + \mathcal{T}_{\tilde{\eta}_{t-1}^s}(\mathrm{grad} f(\tilde{w}^{s-1})))\|_{w_{t-1}^s}^2]
\end{aligned}$$

$$
\begin{aligned}
&\le 2\mathbb{E}_{i_t^s}[\|\mathrm{grad} f_{i_t^s}(w_{t-1}^s) - \mathcal{T}_{\eta_{t-1}^{*s}}(\mathrm{grad} f_{i_t^s}(w^*))\|_{w_{t-1}^s}^2] \\
&\quad + 2\mathbb{E}_{i_t^s}[\|\mathcal{T}_{\tilde{\eta}_{t-1}^s}(\mathrm{grad} f_{i_t^s}(\tilde{w}^{s-1})) - \mathcal{T}_{\eta_{t-1}^{*s}}(\mathrm{grad} f_{i_t^s}(w^*)) - \mathcal{T}_{\tilde{\eta}_{t-1}^s}(\mathrm{grad} f(\tilde{w}^{s-1}))\|_{w_{t-1}^s}^2] \\
&= 2\mathbb{E}_{i_t^s}[\|\mathrm{grad} f_{i_t^s}(w_{t-1}^s) - \mathcal{T}_{\eta_{t-1}^{*s}}(\mathrm{grad} f_{i_t^s}(w^*))\|_{w_{t-1}^s}^2] \\
&\quad + 2\mathbb{E}_{i_t^s}[\|\mathcal{T}_{\tilde{\eta}_{t-1}^s}(\mathrm{grad} f_{i_t^s}(\tilde{w}^{s-1})) - \mathcal{T}_{\eta_{t-1}^{*s}}(\mathrm{grad} f_{i_t^s}(w^*))\|_{w_{t-1}^s}^2] \\
&\quad - 4\langle \mathcal{T}_{\tilde{\eta}_{t-1}^s}(\mathrm{grad} f(\tilde{w}^{s-1})), \mathcal{T}_{\tilde{\eta}_{t-1}^s}(\mathrm{grad} f(\tilde{w}^{s-1})) - \mathcal{T}_{\eta_{t-1}^{*s}}(\mathrm{grad} f(w^*))\rangle_{w_{t-1}^s} \\
&\quad + 2\|\mathrm{grad} f(\tilde{w}^{s-1})\|_{\tilde{w}^{s-1}}^2 \\
&= 2\mathbb{E}_{i_t^s}[\|\mathrm{grad} f_{i_t^s}(w_{t-1}^s) - P^{w_{t-1}^s \leftarrow w^*}(\mathrm{grad} f_{i_t^s}(w^*)) \\
&\qquad\qquad + P^{w_{t-1}^s \leftarrow w^*}(\mathrm{grad} f_{i_t^s}(w^*)) - \mathcal{T}_{\eta_{t-1}^{*s}}(\mathrm{grad} f_{i_t^s}(w^*))\|_{w_{t-1}^s}^2] \\
&\quad + 2\mathbb{E}_{i_t^s}[\|\mathcal{T}_{\tilde{\eta}_{t-1}^s}(\mathrm{grad} f_{i_t^s}(\tilde{w}^{s-1})) - \mathrm{grad} f_{i_t^s}(w_{t-1}^s) \\
&\qquad\qquad + \mathrm{grad} f_{i_t^s}(w_{t-1}^s) - \mathcal{T}_{\eta_{t-1}^{*s}}(\mathrm{grad} f_{i_t^s}(w^*))\|_{w_{t-1}^s}^2] \\
&\quad - 2\|\mathrm{grad} f(\tilde{w}^{s-1})\|_{w^{s-1}}^2 \\
&\le 4\mathbb{E}_{i_t^s}[\|\mathrm{grad} f_{i_t^s}(w_{t-1}^s) - P^{w_{t-1}^s \leftarrow w^*}(\mathrm{grad} f_{i_t^s}(w^*))\|_{w_{t-1}^s}^2] \\
&\quad + 4\mathbb{E}_{i_t^s}[\|P^{w_{t-1}^s \leftarrow w^*}(\mathrm{grad} f_{i_t^s}(w^*)) - \mathcal{T}_{\eta_{t-1}^{*s}}(\mathrm{grad} f_{i_t^s}(w^*))\|_{w_{t-1}^s}^2] \\
&\quad + 4\mathbb{E}_{i_t^s}[\|\mathcal{T}_{\tilde{\eta}_{t-1}^s}(\mathrm{grad} f_{i_t^s}(\tilde{w}^{s-1})) - \mathrm{grad} f_{i_t^s}(w_{t-1}^s)\|_{w_{t-1}^s}^2] \\
&\quad + 4\mathbb{E}_{i_t^s}[\|\mathrm{grad} f_{i_t^s}(w_{t-1}^s) - \mathcal{T}_{\eta_{t-1}^{*s}}(\mathrm{grad} f_{i_t^s}(w^*))\|_{w_{t-1}^s}^2] \\
&\le 4\beta^2(\mathrm{dist}(w_{t-1}^s, w^*))^2 + 4\theta^2\|\eta_{t-1}^{*s}\|_{w^*}^2 \mathbb{E}_{i_t^s}[\|\mathrm{grad} f_{i_t^s}(w^*)\|_{w^*}^2] \\
&\quad + 4\mathbb{E}_{i_t^s}[\|\mathcal{T}_{\tilde{\eta}_{t-1}^s}(\mathrm{grad} f_{i_t^s}(\tilde{w}^{s-1})) - P^{w_{t-1}^s \leftarrow \tilde{w}^{s-1}}(\mathrm{grad} f_{i_t^s}(\tilde{w}^{s-1})) \\
&\qquad\qquad + P^{w_{t-1}^s \leftarrow \tilde{w}^{s-1}}(\mathrm{grad} f_{i_t^s}(\tilde{w}^{s-1})) - \mathrm{grad} f_{i_t^s}(w_{t-1}^s)\|_{w_{t-1}^s}^2] \\
&\quad + 2(4\beta^2(\mathrm{dist}(w_{t-1}^s, w^*))^2 + 4\theta^2\|\eta_{t-1}^{*s}\|_{w^*}^2 \mathbb{E}_{i_t^s}[\|\mathrm{grad} f_{i_t^s}(w^*)\|_{w^*}^2]) \\
&\le 12(\beta^2(\mathrm{dist}(w_{t-1}^s, w^*))^2 + C^2\theta^2\|\eta_{t-1}^{*s}\|_{w^*}^2) \\
&\quad + 8\theta^2\mathbb{E}_{i_t^s}[\|\tilde{\eta}_{t-1}^s\|_{\tilde{w}^{s-1}}^2 \|\mathrm{grad} f_{i_t^s}(\tilde{w}^{s-1})\|_{\tilde{w}^{s-1}}^2] + 8\beta^2(\mathrm{dist}(\tilde{w}^{s-1}, w_{t-1}^s))^2 \\
&\le 4(\beta^2 + \mu^2 C^2\theta^2)(3(\mathrm{dist}(w_{t-1}^s, w^*))^2 + 2(\mathrm{dist}(\tilde{w}^{s-1}, w_{t-1}^s))^2) \\
&\le 4(\beta^2 + \mu^2 C^2\theta^2)(3(\mathrm{dist}(w_{t-1}^s, w^*))^2 + 2(\mathrm{dist}(\tilde{w}^{s-1}, w^*) + \mathrm{dist}(w^*, w_{t-1}^s))^2) \\
&\le 4(\beta^2 + \mu^2 C^2\theta^2)(7(\mathrm{dist}(w_{t-1}^s, w^*))^2 + 4(\mathrm{dist}(\tilde{w}^{s-1}, w^*))^2),
\end{aligned}
$$

where the relation $\|a+b\|^2 \le 2\|a\|^2 + 2\|b\|^2$ for vectors a and b in a norm space and the triangle inequality are used repeatedly. Note also that $\mathbb{E}_{i_t^s}[\mathrm{grad} f_{i_t^s}(\tilde{w}^{s-1})] = \mathrm{grad} f(\tilde{w}^{s-1})$ and $\mathrm{grad} f(w^*) = 0$ and that $\mathbb{E}_{i_t^s}$ is a linear operator. Furthermore, we have evaluated the value $\mathbb{E}_{i_t^s}[\|\mathrm{grad} f_{i_t^s}(w_{t-1}^s) - \mathcal{T}_{\eta_{t-1}^{*s}}(\mathrm{grad} f_{i_t^s}(w^*))\|_{w_{t-1}^s}^2]$ and again used the obtained relation in the third inequality. ∎

Acknowledgment. The authors would like to thank the anonymous referees for their insightful comments that helped improve the paper significantly.

REFERENCES

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

[2] Z. Allen-Zhu and E. Hazan, *Variance reduction for faster non-convex optimization*, in Proceedings of the 33rd International Conference on Machine Learning, Proc. Mach. Learn. Res. 48, 2016, pp. 699–707; available at http://proceedings.mlr.press/v48/.

[3] Z. Allen-Zhu and Y. Yuan, *Improved SVRG for non-strongly-convex or sum-of-non-convex objectives*, in Proceedings of the 33rd International Conference on Machine Learning, Proc. Mach. Learn. Res. 48, 2016, pp. 1080–1089; available at http://proceedings.mlr.press/v48/.
[4] L. Balzano, R. Nowak, and B. Recht, *Online identification and tracking of subspaces from highly incomplete information*, in Proceedings of the 48th Annual Allerton Conference on Communication, Control, and Computing, IEEE Press, Piscataway, NJ, 2010, pp. 704–711.
[5] S. Bonnabel, *Stochastic gradient descent on Riemannian manifolds*, IEEE Trans. Automat. Control, 58 (2013), pp. 2217–2229.
[6] N. Boumal, B. Mishra, P.-A. Absil, and R. Sepulchre, *Manopt: A MATLAB toolbox for optimization on manifolds*, J. Mach. Learn. Res., 15 (2014), pp. 1455–1459.
[7] A. Defazio, F. Bach, and S. Lacoste-Julien, *SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives*, in Adv. Neural Inf. Process. Syst. 27, Curran Associates, Red Hook, NY, 2014, pp. 1646–1654.
[8] D. L. Fisk, *Quasi-martingales*, Trans. Amer. Math. Soc., 120 (1965), pp. 369–389.
[9] D. Garber and E. Hazan, *Fast and Simple PCA via Convex Optimization*, preprint, https://arxiv.org/abs/1509.05647, 2015.
[10] K. Goldberg, T. Roeder, D. Gupta, and C. Perkins, *Eigentaste: A constant time collaborative filtering algorithm*, Inf. Retr., 4 (2001), pp. 133–151.
[11] W. Huang, P.-A. Absil, and K. A. Gallivan, *A Riemannian symmetric rank-one trust-region method*, Math. Program., 150 (2015), pp. 179–216.
[12] W. Huang, K. A. Gallivan, and P.-A. Absil, *A Broyden class of quasi-Newton methods for Riemannian optimization*, SIAM J. Optim., 25 (2015), pp. 1660–1685.
[13] S. Jayasumana, R. Hartley, M. Salzmann, H. Li, and M. Harandi, *Kernel methods on Riemannian manifolds with Gaussian RBF kernels*, IEEE Trans. Pattern Anal. Mach. Intell., 37 (2015), pp. 2464–2477.
[14] B. Jeuris, R. Vandebril, and B. Vandereycken, *A survey and comparison of contemporary algorithms for computing the matrix geometric mean*, Electron. Trans. Numer. Anal., 39 (2012), pp. 379–402.
[15] R. Johnson and T. Zhang, *Accelerating stochastic gradient descent using predictive variance reduction*, in Adv. Neural Inf. Process. Syst. 26, Curran Associates, Red Hook, NY, 2013, pp. 315–323.
[16] H. Kasai, H. Sato, and B. Mishra, *Riemannian Stochastic Variance Reduced Gradient on Grassmann Manifold*, preprint, https://arxiv.org/abs/1605.07367, 2016.
[17] H. Kasai, H. Sato, and B. Mishra, *Riemannian stochastic recursive gradient algorithm*, in Proceedings of the 35th International Conference on Machine Learning, J. Dy and A. Krause, eds., Proc. Mach. Learn. Res. 80, 2018, pp. 2516–2524; available at http://proceedings.mlr.press/v80/.
[18] H. Kasai, H. Sato, and B. Mishra, *Riemannian stochastic quasi-Newton algorithm with variance reduction and its convergence analysis*, in Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics, A. Storkey and F. Perez-Cruz, eds., Proc. Mach. Learn. Res. 84, 2018, pp. 269–278; available at http://proceedings.mlr.press/v84/.
[19] J. Konečný and P. Richtárik, *Semi-Stochastic Gradient Descent Methods*, preprint, https://arxiv.org/abs/1312.1666, 2013.
[20] D. Kressner, M. Steinlechner, and B. Vandereycken, *Low-rank tensor completion by Riemannian optimization*, BIT, 54 (2014), pp. 447–468.
[21] J. Mairal, *Incremental majorization-minimization optimization with application to large-scale machine learning*, SIAM J. Optim., 25 (2015), pp. 829–855.
[22] G. Meyer, S. Bonnabel, and R. Sepulchre, *Linear regression under fixed-rank constraints: A Riemannian approach*, in Proceedings of the 28th International Conference on Machine Learning, 2011, Omnipress, Madison, WI, pp. 545–552.
[23] B. Mishra and R. Sepulchre, *R3MC: A Riemannian three-factor algorithm for low-rank matrix completion*, in Proceedings of the 53rd IEEE Conference on Decision and Control, IEEE Press, Piscataway, NJ, 2014, pp. 1137–1142.
[24] B. Mishra and R. Sepulchre, *Riemannian preconditioning*, SIAM J. Optim., 26 (2016), pp. 635–660.
[25] X. Pennec, P. Fillard, and N. Ayache, *A Riemannian framework for tensor computing*, Int. J. Comput. Vis., 66 (2006), pp. 41–66.
[26] S. J. Reddi, A. Hefny, S. Sra, B. Poczos, and A. Smola, *Stochastic variance reduction for nonconvex optimization*, in Proceedings of the 33rd International Conference on Machine Learning, Proc. Mach. Learn. Res. 48, 2016, pp. 314–323; available at http://proceedings.mlr.press/v48/.

[27] N. L. Roux, M. Schmidt, and F. R. Bach, *A stochastic gradient method with an exponential convergence rate for finite training sets*, in Adv. Neural Inf. Process. Syst. 25, Curran Associates, Red Hook, NY, 2012, pp. 2663–2671.
[28] S. Shalev-Shwartz, *SDCA without Duality*, preprint, https://arxiv.org/abs/1502.06177, 2015.
[29] S. Shalev-Shwartz and T. Zhang, *Proximal Stochastic Dual Coordinate Ascent*, preprint, https://arxiv.org/abs/1211.2717, 2012.
[30] S. Shalev-Shwartz and T. Zhang, *Stochastic dual coordinate ascent methods for regularized loss minimization*, J. Mach. Learn. Res., 14 (2013), pp. 567–599.
[31] O. Shamir, *Fast Stochastic Algorithms for SVD and PCA: Convergence Properties and Convexity*, preprint, https://arxiv.org/abs/1507.08788, 2015.
[32] L. Xiao and Y. Zhang, *A proximal stochastic gradient method with progressive variance reduction*, SIAM J. Optim., 24 (2014), pp. 2057–2075.
[33] X. Yuan, P.-A. Huang, W. Absil, and K. A. Gallivan, *A Riemannian limited-memory BFGS algorithm for computing the matrix geometric mean*, in Proceedings of the International Conference on Computational Science, Procedia Comput. Sci. 80 (2016), pp. 2147–2157.
[34] H. Zhang, S. J. Reddi, and S. Sra, *Riemannian SVRG: Fast stochastic optimization on Riemannian manifolds*, in Adv. Neural Inf. Process. Syst. 29, Curran Associates, Red Hook, NY, 2016, pp. 4592–4600.
[35] H. Zhang and S. Sra, *First-order methods for geodesically convex optimization*, in Conference on Learning Theory 2016, Proc. Mach. Learn. Res. 49, 2016, pp. 1617–1638; available at http://proceedings.mlr.press/v49/.