Vector Transport-Free SVRG with General Retraction for Riemannian Optimization: Complexity Analysis and Practical Implementation

Bo Jiang * Shiqian Ma † Anthony Man-Cho So ‡ Shuzhong Zhang §

May 24, 2017

Abstract

In this paper, we propose a vector transport-free stochastic variance reduced gradient (SVRG) method with general retraction for empirical risk minimization over Riemannian manifold. Existing SVRG methods on manifold usually consider a specific retraction operation, and involve additional computational costs such as parallel transport or vector transport. The vector transport-free SVRG with general retraction we propose in this paper handles general retraction operations, and do not need additional computational costs mentioned above. As a result, we name our algorithm S-SVRG, where the first "S" means simple. We analyze the iteration complexity of S-SVRG for obtaining an ϵ-stationary point and its local linear convergence by assuming the Łojasiewicz inequality, which naturally holds for PCA and holds with high probability for matrix completion problem. We also incorporate the Barzilai-Borwein step size and design a very practical S-SVRG-BB method. Numerical results on PCA and matrix completion problems are reported to demonstrate the efficiency of our methods.

Keywords: Stochastic Variance Reduced Gradient, Riemannian Manifold, Orthogonality Constraints, Principal Component Analysis, Matrix Completion

1 Introduction

The stochastic variance reduced gradient (SVRG) method proposed by Johnson and Zhang [15] has been shown to be very effective for empirical risk minimization problems that involve large-scale training data set in the objective. There have been many variants of SVRG for solving nonsmooth and convex problem [33], nonconvex problem [2, 24, 3], and optimization on manifold [17, 34, 35]. In this paper, we propose a vector transport-free SVRG with general retraction (S-SVRG) that minimizes empirical risk over manifold:

$$\min_{X\in\mathbb{R}^{d\times r}} f(X) \coloneqq \frac{1}{n}\sum_{i=1}^{n} f_i(X), \quad \text{s.t.} \quad X \in \mathcal{M}, \tag{1.1}$$

*School of Mathematical Sciences, Key Laboratory for NSLSCS of Jiangsu Province, Nanjing Normal University, Nanjing 210023, China. Email: jiangbo@njnu.edu.cn.

†Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Shatin, N. T., Hong Kong. Email: sqma@se.cuhk.edu.hk.

‡Department of Systems Engineering and Engineering Management, and, by courtesy, CUHKBGI Innovation Institute of Transomics, The Chinese University of Hong Kong, Shatin, N. T., Hong Kong. Email: manchoso@se.cuhk.edu.hk.

§Department of Industrial and Systems Engineering, University of Minnesota, Minneapolis, MN 55455, USA. Email: zhangs@umn.edu.

where $\mathcal{M}$ denotes a Riemannian manifold, and $f_i(X)\colon \mathcal{M} \to \mathbb{R}$ is differentiable. Although [17, 34, 35, 3] also studied SVRG for solving (1.1), we will discuss later that our S-SVRG is more general and efficient, in the sense that it can handle general objective function, general retraction operation, and does not need additional computational costs such as parallel transport and vector transport that are required in [17, 34, 35].

For the ease of presentation, we mainly focus on discussing the case when the manifold in (1.1) is the set of orthonormal matrices:

$$\min_{X\in\mathbb{R}^{d\times r}} f(X) \coloneqq \frac{1}{n}\sum_{i=1}^{n} f_i(X), \quad \text{s.t.} \quad X \in \mathsf{St}_{d,r} \coloneqq \{X \in \mathbb{R}^{d\times r}\colon X^{\mathsf{T}}X = I_r\}, \tag{1.2}$$

where $f_i(X)\colon \mathsf{St}_{d,r} \to \mathbb{R}$ is differentiable, I_r is the r-th order identity matrix, and $\mathsf{St}_{d,r}$ is the compact Stiefel manifold. The Grassmann manifold $\mathsf{Gr}_{d,r}$ takes the quotient representation as $\mathsf{St}_{d,r}/\mathsf{St}_r$ [10], here $\mathsf{St}_r \coloneqq \mathsf{St}_{r,r}$. Problem (1.2) has wide applications such as principal component analysis (PCA) [16], the Karcher mean problem [17], the joint diagonalization problem [30], the domain adaptation problem [22], and a recovering problem in dictionary learning [28], just to name a few.

1.1 Related works

There are mainly two classes of SVRG methods for manifold optimization. [17, 34, 35] belong to the first class and they construct the stochastic variance reduced Riemannian gradients by invoking parallel transport or vector transport. Specifically, Kasai, Sato and Mishra [17] mainly consider SVRG on Grassmann manifold, and the retraction considered is exponential map [1]. Moreover, the algorithm in [17] requires diminishing step size to guarantee the global convergence and its local linear convergence requires the positiveness of the Riemannian Hessian at a non-degenerate local minimizer. Xu and Ke [34] consider SVRG for eigenvalue problem whose objective function is quadratic and the retraction used is polar decomposition. The linear convergence of the algorithm in [34] requires that the initial point is sufficiently close to the optimal solution. Zhang, Reddi and Sra [35] consider SVRG for general Riemannian manifold (R-SVRG), and the retraction used is the exponential map. However, R-SVRG restricts the objective function f on a compact set $\mathcal{X}$ of the considered manifold. [35] proves that the IFO-calls complexity (i.e., the total number of component gradient evaluations) of R-SVRG for achieving an ϵ-stationary point is $O(\zeta^{\frac{1}{2}}n^{2/3}/\epsilon + n)$, where ζ depends on the sectional curvature and the diameter of $\mathcal{X}$, and this result also requires that f is geodesically L-smooth and the sectional curvature of $\mathcal{X}$ has finite lower and upper bounds. Linear convergence of R-SVRG is shown under the condition that f is globally gradient dominated. Besides, [35] also shows the linear convergence of R-SVRG under the assumption that f is geodesically convex. However, it should be noted that every geodesically convex function on a compact manifold is a constant [5]. Papers in the second class include [26, 25, 32, 3] and they do not need parallel transport or vector transport. Shamir proposes the VR-PCA algorithm [26, 25] for solving PCA problem whose objective function is quadratic. Linear convergence is established under the assumption that the initial point is sufficiently close to the optimal solution. The Stiefel-SVRG proposed by Wu [32] is also for solving PCA problem, and linear convergence to stationary point is shown by proving that the Łojasiewicz inequality holds for PCA problem. Aravkin and Davis [3] propose a very general SVRG algorithm for solving nonconvex and nonsmooth composite problems that include (1.2) as a special case. The IFO-calls complexity of algorithm in [3] is also shown to be $O(n^{2/3}/\epsilon + n)$. However, the retraction operation considered in [3] is gradient projection,

[1] For general Riemannian manifold, it is stated in [1] that "computing the exponential is, in general, a computationally daunting task". For a comparison of the computational cost for the exponential map in $\mathsf{St}_{d,r}$ and other retractions, see [14, 11].

and it is not clear how to extend the result to other retractions. The differences of the existing methods and our S-SVRG are summarized in Table 1.

Table 1: Comparison of SVRG for manifold optimization: "IFO" denotes whether the IFO-calls complexity is analyzed

method	obj.	manifold	retraction	additional cost	IFO
VR-PCA [25]	quadratic	Stiefel	gradient projection	—	×
Stiefel-SVRG [32]	quadratic	Stiefel	general	—	×
R-SVRG [17]	general	Grassmann	exponential	parallel translation	×
SVRRG [34]	quadratic	Stiefel	polar decomposition	vector transport	×
RSVRG [35]	general	compact	exponential	parallel translation	√
SMART-SVRG [3]	general	general	gradient projection	—	√
Our S-SVRG	general	general	general	—	√

1.2 Our contributions

Our main contributions lie in several folds. (i) Our S-SVRG handles general objective function, general manifold, and general retraction steps (including all the seven retractions that will be discussed in Appendix A), and it does not need additional costs such as parallel transport or vector transport. (ii) We analyze the convergence and IFO-calls complexity of S-SVRG, and we do not need the various conditions in [35]. As a by-product, we also analyze the convergence and IFO-calls complexity of a vector transport-free randomized stochastic gradient descent with general retraction (S-SGD) for solving (1.2). (iii) We prove the local linear convergence of S-SVRG using Łojasiewicz inequality, which holds naturally for PCA problem and holds with high probability for matrix completion problems. Therefore, the condition required for local linear convergence is weaker than the ones used in the previous works [17, 34, 35, 3]. (iv) We incorporate the Barzilai-Borwein step size to S-SVRG and design a very practical S-SVRG-BB method, and this resolves the issue of choosing step size in S-SVRG.

1.3 Organization

The rest of this paper is organized as follows. We introduce some preliminaries on manifold optimization in Section 2. We propose our S-SVRG method for solving (1.2) and analyze its IFO-calls complexity and local linear convergence in Section 3. As a by-product, we also provide an analysis of SGD for solving (1.2). Extensions to more general manifolds are studied in Section 4. We propose the S-SVRG-BB method in Section 5. Numerical results on PCA and matrix completion problems are presented in Section 6. Finally, we conclude the paper in Section 7.

2 Preliminaries

Throughout this paper, we make the following assumption for (1.2).

Assumption 2.1. *$f_i(X)\colon \mathsf{St}_{d,r} \to \mathbb{R}$ is differentiable, and its Euclidean gradient $\nabla f_i(X)$ is L-Lipschitz continuous over $\mathsf{St}_{d,r}$, i.e.,*

$$\|\nabla f_i(X) - \nabla f_i(Y)\|_{\mathsf{F}} \le L\|X - Y\|_{\mathsf{F}}, \quad \forall X, Y \in \mathsf{St}_{d,r}. \tag{2.1}$$

It follows that $\nabla f(X)$ is also L-Lipschitz continuous over $\mathsf{St}_{d,r}$.

We first introduce some basic notions of optimization on Riemannian manifold $\mathcal{M}$. For each $X \in \mathcal{M}$, the tangent space is denoted by $\mathbf{T}_X\mathcal{M}$. We define the *inner product* on $\mathbf{T}_X\mathcal{M}$ as $\langle\cdot,\cdot\rangle_X$ and the corresponding induced norm $\|E\|_X = \sqrt{\langle E, E\rangle_X}$ is an equivalent norm to the Frobenius norm, namely, there exist $\nu, \gamma > 0$ such that

$$\nu\|E\|_{\mathsf{F}}^2 \le \langle E, E\rangle_X \le \gamma\|E\|_{\mathsf{F}}^2, \quad \forall E \in \mathbf{T}_X\mathcal{M}. \tag{2.2}$$

The Riemannian gradient $\operatorname{grad} f(X)$ is the unique element of $\mathbf{T}_X\mathcal{M}$ satisfying

$$\langle \operatorname{grad} f(X), E\rangle_X = \mathrm{D}f(X)[E] = \langle \nabla f(X), E\rangle, \quad \forall E \in \mathbf{T}_X\mathcal{M}, \tag{2.3}$$

where $\langle\cdot,\cdot\rangle$ is the Euclidean inner product.

The feasible gradient descent method is based on the notion of retraction. Given any $X \in \mathcal{M}$, the *retraction* $\mathcal{R}(t) \coloneqq \mathcal{R}(X, tE)$ along the direction $E \in \mathbf{T}_X\mathcal{M}$ is a smooth map from $\mathbf{T}_X\mathcal{M}$ to $\mathcal{M}$ that satisfies

$$\mathcal{R}(0) = X, \ \mathcal{R}'(0) = E, \quad \forall t \in [0, T_{\mathcal{R}}], \tag{2.4}$$

where $\mathcal{R}'(0) = \frac{\mathrm{d}}{\mathrm{d}t}\mathcal{R}(t)\big|_{t=0}$. Starting from a given initial point X^0, the feasible method based on the retraction updates the iterates as

$$X^{k+1} = \mathcal{R}\big(X^k, \tau_k E^k\big), \ k = 0, 1, 2, \dots \tag{2.5}$$

where $\tau_k > 0$ is the step size and $E^k \in \mathbf{T}_X\mathcal{M}$.

In the following we give the definitions of stationary point and ϵ-stationary point of (1.1).

Definition 2.2 (Stationary point)**.** *$X \in \mathcal{M}$ is called a stationary point of problem* (1.1) *if* $\operatorname{grad} f(X) = 0$.

Definition 2.3 (ϵ-stationary point)**.** *$X \in \mathcal{M}$ is called an ϵ-stationary point of problem* (1.1) *if* $\|\operatorname{grad} f(X)\|_{\mathsf{F}}^2 \le \epsilon$.

Definition 2.4 (stochastic ϵ-stationary point)**.** *Suppose $X^r \in \mathcal{M}$ is returned by a stochastic feasible method for* (1.1)*, we call X^r a stochastic ϵ-stationary point if*

$$\mathbb{E}\big[\|\operatorname{grad} f(X^r)\|_{\mathsf{F}}^2\big] \le \epsilon,$$

where the expectation is taken with respect to the whole stochasticity of the algorithm.

With slight abuse of notation, we still use $\operatorname{grad} f(X)$ to denote Riemannian gradient on $\mathsf{St}_{d,r}$. We now give a complete description of $\operatorname{grad} f(X)$. For each $X \in \mathsf{St}_{d,r}$, the tangent space at X is referred to $\mathbf{T}_X\mathsf{St}_{d,r} \coloneqq \{Z \in \mathbb{R}^{d\times r} \colon X^{\mathsf{T}}Z + Z^{\mathsf{T}}X = 0\}$. For any $\rho > 0$, define the inner product on $\mathbf{T}_X\mathsf{St}_{d,r}$ as $\langle E_1, E_2\rangle_X = \langle E_1, \mathbf{P}_{\rho,X}E_2\rangle, \forall E_1, E_2 \in \mathbf{T}_X\mathsf{St}_{d,r}$, where $\mathbf{P}_{\rho,X} = I_d - \big(1 - 1/(4\rho)\big)XX^{\mathsf{T}}$. Each point in $\mathsf{Gr}_{d,r}$ is essential an equivalent class $[X] = \{XQ_r \colon Q_r \in \mathsf{St}_r\}$, where St_r stands for $\mathcal{O}_{r,r}$ for short. The tangent space at $[X]$ is given as $\mathbf{T}_{[X]}\mathsf{Gr}_{d,r} = \{Z \in \mathbb{R}^{d\times r} \colon X^{\mathsf{T}}Z = 0\}$ and the corresponding inner product is always taken as $\langle E_1, E_2\rangle_{[X]} = \langle E_1, E_2\rangle, \ \forall E_1, E_2 \in \mathbf{T}_{[X]}\mathsf{Gr}_{d,r}$; see [10]. For simplicity of notation, we denote $[X]$ by X. Given $\rho \ge 0$, define the operator

$$\mathbf{D}_\rho(X, Y) \coloneqq (I_r - XX^{\mathsf{T}})Y + 4\rho X\operatorname{skew}(X^{\mathsf{T}}Y), \tag{2.6}$$

where $\operatorname{skew}(X^{\mathsf{T}}Y) = (X^{\mathsf{T}}Y - Y^{\mathsf{T}}X)/2$. From (2.3), we know the Riemannian gradient on $\mathsf{St}_{d,r}$ can be defined as

$$\operatorname{grad} f(X) = \mathbf{D}_\rho(X, \nabla f(X))$$

with $\rho = 0$ for $\mathsf{Gr}_{d,r}$ and $\rho > 0$ for $\mathsf{St}_{d,r}$. Note that when $X^{\mathsf{T}}\nabla f(X) = \nabla f(X)^{\mathsf{T}}X$, we have $\operatorname{grad} f(X) \equiv \mathbf{D}_0(X, \nabla f(X))$. Besides, (2.2) holds for $\mathsf{St}_{d,r}$ with $\nu = \min\{1, 1/(4\rho)\}$ and $\gamma = 1$.

3 A S-SVRG method for problem (1.2)

In this section we propose a S-SVRG method for solving (1.2). We first establish a sufficient decrease property of retraction of $\mathsf{St}_{d,r}$, which plays an important role in establishing the complexity results of S-SVRG.

3.1 Sufficient decrease property

Consider a retraction $\mathcal{R}(t)$ of $\mathsf{St}_{d,r}$ with $\mathcal{R}(0) = X$. [2] We aim to establish the so-called sufficient decrease property of $f(\mathcal{R}(t))$, i.e., there is a sufficient reduction from $f(X)$ to $f(\mathcal{R}(t))$. By the compactness of $\mathsf{St}_{d,r}$, there exists a constant $C > 0$ such that

$$\|\nabla f(X)\|_{\mathsf{F}} \le C, \quad \forall X \in \mathsf{St}_{d,r}.$$

Following the proof of Lemma 3 in [7], we know that $\mathcal{R}(t)$ satisfies the following properties.

Proposition 3.1. *There exist constants $L_1 \ge 1$, $L_2 > 0$ such that*

$$\|\mathcal{R}(t) - \mathcal{R}(0)\|_{\mathsf{F}} \le L_1 t \|\mathcal{R}'(0)\|_{\mathsf{F}}, \tag{3.1}$$

$$\|\mathcal{R}(t) - \mathcal{R}(0) - t\mathcal{R}'(0)\|_{\mathsf{F}} \le L_2 t^2 \|\mathcal{R}'(0)\|_{\mathsf{F}}^2, \tag{3.2}$$

hold for any $t \ge 0$.

We have the following remarks regarding Proposition 3.1. For the particular case $\mathcal{M} = \mathbb{R}^{n\times r}$ and $\mathcal{R}(t) = X + tE$, we always have $L_1 = 1$ and $L_2 = 0$, for given $X, E \in \mathbb{R}^{n\times r}$. For retractions on $\mathsf{St}_{d,r}$, we can compute L_1 and L_2 explicitly. For example, when $\mathcal{R}(t)$ is polar decomposition, we have $L_1 = 1$, $L_2 = 1/2$, and when $\mathcal{R}(t)$ is QR factorization, we have $L_1 = 1 + \sqrt{2}/2$, $L_2 = \sqrt{10}/2$. Note that these estimations of L_1 and L_2 are much better than those in [9, 32]. The corresponding proofs are given in Appendix A.3.

Second, for any $Z \in \mathbb{R}^{d\times r}$, there holds that

$$\begin{aligned}\langle Z, \mathcal{R}(t) - \mathcal{R}(0)\rangle &= t\left\langle Z, \mathcal{R}'(0)\right\rangle + \left\langle Z, \mathcal{R}(t) - \mathcal{R}(0) - t\mathcal{R}'(0)\right\rangle \\ &\le t\left\langle Z, \mathcal{R}'(0)\right\rangle + \|Z\|_{\mathsf{F}}\|\mathcal{R}(t) - \mathcal{R}(0) - t\mathcal{R}'(0)\|_{\mathsf{F}} \\ &\le t\left\langle Z, \mathcal{R}'(0)\right\rangle + L_2 t^2 \|Z\|_{\mathsf{F}}\|\mathcal{R}'(0)\|_{\mathsf{F}}^2, \end{aligned} \tag{3.3}$$

where the second inequality is due to (3.2). Inequality (3.3) will be used later in our analysis.

We are now ready to present the sufficient decrease property, whose proof can be found in Lemma 3 of [7]. For completeness, we give a simple proof here.

Lemma 3.2. *For any $t \ge 0$, there holds that*

$$f(\mathcal{R}(t)) \le f(X) + t\langle \operatorname{grad} f(X), \mathcal{R}'(0)\rangle_X + \frac{\hat{L}}{2} t^2 \|\mathcal{R}'(0)\|_{\mathsf{F}}^2, \tag{3.4}$$

where $\hat{L} = 2L_2 C + L_1^2 L$.

Proof. From (2.1), we know that $\nabla f(X)$ is L-Lipschitz continuous. Using $\mathcal{R}(0) = X$, we have

$$f(\mathcal{R}(t)) \le f(X) + \langle \nabla f(X), \mathcal{R}(t) - \mathcal{R}(0)\rangle + \frac{L}{2}\|\mathcal{R}(t) - \mathcal{R}(0)\|_{\mathsf{F}}^2. \tag{3.5}$$

It follows from (2.3) that $\langle \nabla f(X), \mathcal{R}'(0)\rangle = \langle \operatorname{grad} f(X), \mathcal{R}'(0)\rangle_X$, which together with (3.3) and (3.1) implies (3.4). □

It is shown in Lemma 3 of [7] that if $\mathcal{M}$ is a compact Riemannian submanifold of some Euclidean space and f has Lipschitz continuous gradient in the convex hull of $\mathcal{M}$, then (3.4) must hold because (3.1) and (3.2) hold. However, if $\mathcal{M}$ is not compact, it remains unknown whether (3.4) holds for some universal $T_{\mathcal{R}}$ and $t \in [0, T_{\mathcal{R}}]$.

[2] For the retractions considered in this paper, we know that $T_{\mathcal{R}} = +\infty$; see Appendices A.1 and A.2.

3.2 A S-SVRG method

Our S-SVRG method is described in Algorithm 1. The random event in Line 6 of Algorithm 1 is denoted by $\xi_{s,k}$. Clearly, $\xi_{s,k}$ is mutually independent of each other. For fixed s, we simply denote $X^{s,k}$, $\xi_{s,k}$, τ_s and $\mathcal{G}^{\mathrm{R}}(X^{s,k},\xi_{s,k})$ respectively by X^k, ξ_k, τ and $\mathcal{G}^{\mathrm{R}}(X^k,\xi_k)$. With the stochastic *Euclidean* gradient $\mathcal{G}(X^k,\xi_k)$ in hand, we can compute the stochastic *Riemannian* gradient $\mathcal{G}^{\mathrm{R}}(X^k,\xi_k)$ such that

$$\langle \mathcal{G}^{\mathrm{R}}(X^k,\xi_k), E\rangle_{X^k} = \langle \mathcal{G}(X^k,\xi_k), E\rangle, \quad \forall E \in \mathbf{T}_{X^k}\mathsf{St}_{d,r}, \tag{3.6}$$

which gives $\mathcal{G}^{\mathrm{R}}(X^k,\xi_k) = \mathbf{D}_\rho\left(X^k, \mathcal{G}(X^k,\xi_k)\right)$, where the operator $\mathbf{D}_\rho(\cdot,\cdot)$ is defined in (2.6).

Algorithm 1 A S-SVRG method for problem (1.2)

1: Given $X^{0,0} \in \mathsf{St}_{d,r}$, the retraction type $\mathcal{R}(\cdot,\cdot)$, the direction parameter $\rho \geq 0$ in $\mathbf{D}_\rho$.
2: Choose the maximal inner iteration number $K \geq 1$ and the mini-batch size $|\mathsf{B}| \geq 1$.
3: **for** $s = 0,\ldots,S-1$ **do**
4: Compute the full *Euclidean* gradient $\nabla f(X^{s,0})$ and set the step size $\tau_s > 0$.
5: **for** $k = 0,\ldots,K-1$ **do**
6: Generate a uniformly random sample $\mathsf{B} \subseteq \{1,\ldots,n\}$ with replacement.
7: Compute the stochastic *Euclidean* gradient $\mathcal{G}(X^{s,k},\xi_{s,k})$ as

$$\mathcal{G}(X^{s,k},\xi_{s,k}) = \nabla f(X^{s,0}) + \frac{1}{|\mathsf{B}|}\sum_{i\in\mathsf{B}}\left(\nabla f_i(X^{s,k}) - \nabla f_i(X^{s,0})\right). \tag{3.7}$$

8: Compute the stochastic *Riemannian* gradient as [3]

$$\mathcal{G}^{\mathrm{R}}(X^{s,k},\xi_{s,k}) = \mathbf{D}_\rho\big(X^{s,k}, \mathcal{G}(X^{s,k},\xi_{s,k})\big). \tag{3.8}$$

9: Update $X^{s,k+1}$ along the direction $-\mathcal{G}^{\mathrm{R}}(X^{s,k},\xi_{s,k})$, i.e.,

$$X^{s,k+1} = \mathcal{R}\big(X^{s,k}, -\tau_s\mathcal{G}^{\mathrm{R}}(X^{s,k},\xi_{s,k})\big). \tag{3.9}$$

10: **end for**
11: Set $X^{s+1,0} \coloneqq X^{s,K}$. Set X_r^s to be $X^{s,k}$ with probability $p_{s,k}, k = 0,\ldots,K$.
12: **end for**
13: **return** X_r uniformly from $\{X_r^s\}$ with $s \in \{0,1,\ldots,S-1\}$.

We now show that the stochastic Riemannian gradient $\mathcal{G}^{\mathrm{R}}(X^k,\xi_k)$ is unbiased and its variance can be well controlled.

Lemma 3.3. *For the sequences generated by Algorithm 1, it holds that*

$$\mathbb{E}_{\xi_k}\left[\mathcal{G}^{\mathrm{R}}(X^k,\xi_k) - \operatorname{grad} f(X^k)\right] = 0, \tag{3.10}$$

$$\mathbb{E}_{\xi_k}\left[\|\mathcal{G}^{\mathrm{R}}(X^k,\xi_k) - \operatorname{grad} f(X^k)\|_{\mathsf{F}}^2\right] \leq \frac{L^2}{\nu^2|\mathsf{B}|}\|X^k - X^0\|_{\mathsf{F}}^2, \tag{3.11}$$

where the constant ν, defined in (2.2)*, equals to* $\min\{1, 1/(4\rho)\}$.

Proof. It is easy to see that $\mathbb{E}_{\xi_k}\left[\mathcal{G}(X^k,\xi_k) - \nabla f(X^k)\right] = 0$. Taking the expectation over ξ_k on both sides of (3.6), we have

$$\langle \mathbb{E}_{\xi_k}[\mathcal{G}^{\mathrm{R}}(X^k,\xi_k)], E\rangle_{X^k} = \langle \nabla f(X^k), E\rangle, \quad \forall E \in \mathbf{T}_{X^k}\mathsf{St}_{d,r}, \tag{3.12}$$

[3] For some special retractions, such as the gradient projection and gradient reflection retractions on $\mathsf{St}_{d,r}$, it is not necessary to compute $\mathcal{G}^{\mathrm{R}}(X^{s,k},\xi_{s,k})$ explicitly.

which together with (2.3) and the uniqueness of grad $f(X^k)$ implies (3.10).

We now prove (3.11). By (3.6) and (2.3), we have

$$\langle \mathcal{G}^{\mathrm{R}}(X^k,\xi_k) - \operatorname{grad} f(X^k), E\rangle_{X^k} = \langle \mathcal{G}(X^k,\xi_k) - \nabla f(X^k), E\rangle, \quad \forall E \in \mathbf{T}_{X^k}\mathsf{St}_{d,r}. \tag{3.13}$$

Letting $E = \mathcal{G}^{\mathrm{R}}(X^k,\xi_k) - \operatorname{grad} f(X^k)$ in (3.13), we have from (2.2) that

$$\nu\|\mathcal{G}^{\mathrm{R}}(X^k,\xi_k) - \operatorname{grad} f(X^k)\|_{\mathsf{F}}^2 \le \|\mathcal{G}(X^k,\xi_k) - \nabla f(X^k)\|_{\mathsf{F}} \cdot \|\mathcal{G}^{\mathrm{R}}(X^k,\xi_k) - \operatorname{grad} f(X^k)\|_{\mathsf{F}},$$

which yields

$$\|\mathcal{G}^{\mathrm{R}}(X^k,\xi_k) - \operatorname{grad} f(X^k)\|_{\mathsf{F}}^2 \le \frac{1}{\nu^2}\|\mathcal{G}(X^k,\xi_k) - \nabla f(X^k)\|_{\mathsf{F}}^2. \tag{3.14}$$

Taking the expectation over ξ_k on both sides of (3.14) leads to

$$\mathbb{E}_{\xi_k}\big[\|\mathcal{G}^{\mathrm{R}}(X^k,\xi_k) - \operatorname{grad} f(X^k)\|_{\mathsf{F}}^2\big] \le \frac{1}{\nu^2}\cdot \mathbb{E}_{\xi_k}[\|\mathcal{G}(X^k,\xi_k) - \nabla f(X^k)\|_{\mathsf{F}}^2]. \tag{3.15}$$

On the other hand, we have

$$\mathbb{E}_{\xi_k}\big[\|\mathcal{G}(X^k,\xi_k) - \nabla f(X^k)\|_{\mathsf{F}}^2\big] \le \frac{1}{n|\mathsf{B}|}\frac{n-|\mathsf{B}|}{n-1}\sum_{i=1}^{n}\|\nabla f_i(X^k) - \nabla f_i(X^0)\|_{\mathsf{F}}^2 \le \frac{L^2}{|\mathsf{B}|}\|X^k - X^0\|_{\mathsf{F}}^2,$$

where the first inequality comes from Lemma 4 in [19], and the second one is due to (2.1). Combining the above inequality and (3.15), we have (3.11). □

The following lemma plays a key role in establishing the iteration complexity. Its proof is relegated to Appendix B.

Lemma 3.4. *Consider the sequences* $\{\mathbf{a}_k \ge 0\}, \{\mathbf{b}_k \ge 0\}, \{\mathbf{f}_k\}$ *with* $k = 0, \ldots, K$ *and* $\mathbf{b}_0 = 0$. *If there exist positive constants* $\mathbf{a}, \mathbf{b} \ne 1, \mathbf{c}, \mathbf{d}$ *such that*

$$\mathbf{f}_{k+1} \le \mathbf{f}_k - \mathbf{c}\mathbf{a}_k + \mathbf{d}\mathbf{b}_k, \tag{3.16}$$

$$\mathbf{b}_{k+1} \le \mathbf{b}\mathbf{b}_k + \mathbf{a}\mathbf{a}_k \tag{3.17}$$

hold for $k = 0, \ldots, K-1$. *Then we have*

$$\mathbf{f}_K \le \mathbf{f}_0 - \sum_{k=0}^{K-1}\Delta_k\mathbf{a}_k \quad \textit{with} \quad \Delta_k = \mathbf{c} - \mathbf{a}\mathbf{d}\Gamma(b, K-k), \tag{3.18}$$

where the function $\Gamma(\cdot,\cdot)$ *is defined as* $\Gamma(z,i) = \frac{(1+z)^{i-1}-1}{z}$.

3.3 Iteration complexity of S-SVRG

We first show that the function value over one epoch, i.e., one outer loop, has sufficient reduction in expectation.

Lemma 3.5. *Consider Algorithm 1. For fixed s, it holds that*

$$\mathbb{E}_{\xi_{[K-1]}}[f(X^K)] \le f(X^0) - \sum_{k=0}^{K-1}\Delta_k\mathbb{E}_{\xi_{[K-1]}}\big[\|\operatorname{grad} f(X^k)\|_{\mathsf{F}}^2\big], \tag{3.19}$$

where $\xi_{[K-1]} = (\xi_0, \ldots, \xi_{K-1})$ *and*

$$\frac{\Delta_k}{\tau} = \nu - \frac{\hat{L}\tau}{2}\left[1 + \left(1 + \frac{2}{\tilde{L}\beta\tau}\right)\frac{\tilde{L}L^2\tau^2}{\nu^2|\mathsf{B}|}\Gamma_k\right], \tag{3.20}$$

where $\beta > 0$ *is a constant,* $\tilde{L} = L_1^2 + 4L_2\sqrt{r}$ *and*

$$\Gamma_k = \Gamma\left(2\beta\tau + \frac{\tilde{L}L^2\tau^2}{\nu^2|\mathsf{B}|}, K-k\right). \tag{3.21}$$

Proof. The proof consists of three steps.

First, we establish the sufficient reduction of the function value per inner iteration. By (3.10), we have

$$\mathbb{E}_{\xi_k}\big[\|\mathcal{G}^{\mathrm{R}}(X^k,\xi_k)\|_{\mathsf{F}}^2\big] = \mathbb{E}_{\xi_k}\big[\|\mathcal{G}^{\mathrm{R}}(X^k,\xi_k) - \operatorname{grad} f(X^k)\|_{\mathsf{F}}^2\big] + \|\operatorname{grad} f(X^k)\|_{\mathsf{F}}^2. \tag{3.22}$$

Letting $\mathcal{R}'(0) = -\mathcal{G}^{\mathrm{R}}(X^k,\xi_k)$ in (3.4), and taking the expectation over ξ_k on both sides of the resulting inequality, we have from $X^k = \mathcal{R}(X^k,0)$, (3.9) and (3.22) that

$$\begin{aligned}\mathbb{E}_{\xi_k}[f(X^{k+1})] \leq f(X^k) - \tau\big\langle \operatorname{grad} f(X^k), \operatorname{grad} f(X^k)\big\rangle_{X^k} + \frac{1}{2}\hat{L}\tau^2\|\operatorname{grad} f(X^k)\|_{\mathsf{F}}^2 \\ + \frac{1}{2}\hat{L}\tau^2\mathbb{E}_{\xi_k}\big[\|\mathcal{G}^{\mathrm{R}}(X^k,\xi_k) - \operatorname{grad} f(X^k)\|_{\mathsf{F}}^2\big].\end{aligned} \tag{3.23}$$

Plugging (3.11) into (3.23) and using (2.2) with $E = \operatorname{grad} f(X^k)$, we obtain

$$\mathbb{E}_{\xi_k}[f(X^{k+1})] \leq f(X^k) - \frac{\tau}{2}\left(2\nu - \hat{L}\tau\right)\|\operatorname{grad} f(X^k)\|_{\mathsf{F}}^2 + \frac{\hat{L}L^2\tau^2}{2\nu^2|\mathsf{B}|}\|X^k - X^0\|_{\mathsf{F}}^2. \tag{3.24}$$

Second, we prove the following inequality:

$$\begin{aligned}&\mathbb{E}_{\xi_k}\big[\|X^{k+1} - X^0\|_{\mathsf{F}}^2\big] \\ \leq\ &\left(1 + 2\beta\tau + \frac{\tilde{L}L^2}{\nu^2|\mathsf{B}|}\tau^2\right)\|X^k - X^0\|_{\mathsf{F}}^2 + \tau\left(\tilde{L}\tau + \frac{2}{\beta}\right)\|\operatorname{grad} f(X^k)\|_{\mathsf{F}}^2.\end{aligned} \tag{3.25}$$

To prove this, we first obtain from $X^k = \mathcal{R}(X^k,0)$ and (3.9) that

$$\begin{aligned}\|X^{k+1} - X^0\|_{\mathsf{F}}^2 = \|X^k - X^0\|_{\mathsf{F}}^2 + \|\mathcal{R}\big(X^k, -\tau\mathcal{G}^{\mathrm{R}}(X^k,\xi_k)\big) - \mathcal{R}(X^k,0)\|_{\mathsf{F}}^2 \\ + 2\big\langle X^k - X^0, \mathcal{R}(X^k, -\tau\mathcal{G}^{\mathrm{R}}(X^k,\xi_k)) - \mathcal{R}(X^k,0)\big\rangle.\end{aligned} \tag{3.26}$$

It follows from (3.3) and $\mathcal{R}'(0) = -\mathcal{G}^{\mathrm{R}}(X^k,\xi_k)$ that

$$\begin{aligned}&\big\langle X^k - X^0, \mathcal{R}(X^k, -\tau\mathcal{G}^{\mathrm{R}}(X^k,\xi_k)) - \mathcal{R}(X^k,0)\big\rangle \\ \leq\ &-\tau\big\langle X^k - X^0, \mathcal{G}^{\mathrm{R}}(X^k,\xi_k)\big\rangle + L_2\tau^2\|X^k - X^0\|_{\mathsf{F}}\|\mathcal{G}^{\mathrm{R}}(X^k,\xi_k)\|_{\mathsf{F}}^2,\end{aligned}$$

which, together with the fact that $\|X^k - X^0\|_{\mathsf{F}} \leq 2\sqrt{r}$, (3.26) and the definition of $\tilde{L}$ implies

$$\|X^{k+1} - X^0\|_{\mathsf{F}}^2 \leq \|X^k - X^0\|_{\mathsf{F}}^2 + \tilde{L}\tau^2\|\mathcal{G}^{\mathrm{R}}(X^k,\xi_k)\|_{\mathsf{F}}^2 - 2\tau\big\langle X^k - X^0, \mathcal{G}^{\mathrm{R}}(X^k,\xi_k)\big\rangle. \tag{3.27}$$

Taking expectation over both sides of (3.27) with respect to ξ_k, and using (3.10), we have

$$\mathbb{E}_{\xi_k}\big[\|X^{k+1} - X^0\|_{\mathsf{F}}^2\big] \leq \|X^k - X^0\|_{\mathsf{F}}^2 + \tilde{L}\tau^2\mathbb{E}_{\xi_k}\big[\|\mathcal{G}^{\mathrm{R}}(X^k,\xi_k)\|_{\mathsf{F}}^2\big] - 2\tau\big\langle X^k - X^0, \operatorname{grad} f(X^k)\big\rangle. \tag{3.28}$$

Combining (3.22) and (3.11), we have

$$\mathbb{E}_{\xi_k}\big[\|\mathcal{G}^{\mathrm{R}}(X^k,\xi_k)\|_{\mathsf{F}}^2\big] \leq \frac{L^2}{\nu^2|\mathsf{B}|}\|X^k - X^0\|_{\mathsf{F}}^2 + \|\operatorname{grad} f(X^k)\|_{\mathsf{F}}^2. \tag{3.29}$$

On the other hand, it follows from the Cauchy-Schwarz inequality that

$$-\big\langle X^k - X^0, \mathcal{G}^{\mathrm{R}}(X^k,\xi_k)\big\rangle \leq \beta\|X^k - X^0\|_{\mathsf{F}}^2 + \frac{1}{\beta}\|\operatorname{grad} f(X^k)\|_{\mathsf{F}}^2. \tag{3.30}$$

Plugging (3.29) and (3.30) into (3.28), we obtain (3.25).

Finally, using (3.24), (3.25) and the fact that ξ_k is independent of each other, considering the sequences $\{\mathbf{a}_k\}, \{\mathbf{b}_k\}, \{\mathbf{f}_k\}$ in Lemma 3.4 as $\{\mathbb{E}_{\xi_{[K-1]}}\big[\|\operatorname{grad} f(X^k)\|_{\mathsf{F}}^2\big]\}$, $\{\mathbb{E}_{\xi_{[K-1]}}\big[\|X^k - X^0\|_{\mathsf{F}}^2\big]\}$ and $\{\mathbb{E}_{\xi_{[K-1]}}[f(X^k)]\}$, respectively, we finally establish (3.19). □

For fixed s, (3.20) implies that $\frac{\tau}{2}\big(2\nu - \hat{L}\tau\big) = \Delta_{K-1} > \cdots > \Delta_0$. Thus

$$\Delta_{\min} := \min_{s,k} \Delta_{s,k} = \min_{s} \Delta_{s,0}.$$

We now bound the variance of grad $f(X_r)$ in expectation.

Lemma 3.6. *Consider Algorithm 1. Suppose that $\Delta_{\min} > 0$. Let $p_{s,k} = \frac{\Delta_{s,k}}{\sum_{k=0}^{K-1}\Delta_{s,k}}$, $k = 0, \ldots, K-1$ and $p_{s,K} = 0$. We have*

$$\mathbb{E}[\|\mathrm{grad}\, f(X_r)\|^2] \le \frac{1}{SK\Delta_{\min}}\big(f(X^{0,0}) - f(X^*)\big), \tag{3.31}$$

where X^* is the optimal solution of problem (1.2), and the expectation is taken over $\xi_{[S-1],[K-1]} = (\xi_{0,[K-1]}, \ldots, \xi_{S-1,[K-1]})$ and the randomness of r.

Proof. First, note that $p_{s,K} = 0$, we have

$$\begin{aligned}\mathbb{E}_{\xi_{s,[K-1]}}[\|\mathrm{grad}\, f(X_r^s)\|_{\mathsf{F}}^2] &= \sum_{k=0}^{K-1} p_{s,k}\, \mathbb{E}_{\xi_{[K-1]}}\big[\|\mathrm{grad}\, f(X^{s,k})\|_{\mathsf{F}}^2\big] \\ &\le \frac{f(X^{s,0}) - \mathbb{E}_{\xi_{[K-1]}}[f(X^{s,K})]}{\sum_{k=0}^{K-1}\Delta_{s,k}} \le \frac{f(X^{s,0}) - \mathbb{E}_{\xi_{[K-1]}}[f(X^{s,K})]}{K\Delta_{\min}},\end{aligned} \tag{3.32}$$

where the first inequality is due to (3.19). Further note that

$$\mathbb{E}[\|\mathrm{grad}\, f(X_r)\|^2] = \frac{1}{S}\sum_{s=0}^{S-1}\mathbb{E}_{\xi_s,[K-1]}[\|\mathrm{grad}\, f(X_r^s)\|_{\mathsf{F}}^2],$$

we know from (3.32) and $X^{s+1,0} = X^{s,K}$ that $\mathbb{E}[\|\mathrm{grad}\, f(X_r)\|^2] \le \frac{1}{SK\Delta_{\min}}\big(f(X^{s,0}) - \mathbb{E}\big[f(X^{S,K})\big]\big)$, which, together with $\mathbb{E}\big[f(X^{S,K})\big] \ge f(X^*)$ implies (3.31). □

By choosing the parameters $K, |\mathsf{B}|, \beta$ and τ_s carefully, we are ready to establish the iteration complexity result of S-SVRG.

Theorem 3.7. *Consider Algorithm 1. Given constants $0 \le \mu \le 2/3$ and $\kappa > 0$, we set the parameters as*

$$K = \lceil (\kappa n)^{\frac{1}{3(1-\mu)}} \rceil, \quad |\mathsf{B}| = \lceil K^{2-3\mu} \rceil, \quad \beta = \frac{\sqrt{\tilde{L}L}}{\nu} K^{\mu-1}, \quad \tau_s \equiv \frac{c\nu}{\sqrt{\tilde{L}L}} K^{-\mu}, \tag{3.33}$$

where the constant $c \in (0,1)$ satisfies

$$\frac{\hat{L}}{\sqrt{\tilde{L}L}} \exp(c^2 + 2c)c \le 1. \tag{3.34}$$

To achieve a stochastic ϵ-stationary point (defined in Definition 2.4) of problem (1.2), the IFO-calls and RO(retraction oracle)-calls complexities are $O(n^{2/3}/\epsilon + n)$ and $O(n^{\frac{\mu}{3(1-\mu)}}/\epsilon + n^{\frac{1}{3(1-\mu)}})$, respectively. In particular, when $\mu = 0$, namely, $|\mathsf{B}| = \lceil (\kappa n)^{2/3} \rceil$, the IFO-calls and RO-calls complexities become $O(n^{2/3}/\epsilon + n)$ and $O(1/\epsilon + n^{1/3})$, respectively.

Proof. Since all τ_s are the same, again we drop the subscript s for simplicity.

We first give the lower bound of $\Delta_{\min}$. Note that $\tilde{L} \ge L_1^2 \ge 1$ and $0 \le \mu \le 2/3$, with the choices of τ and $|\mathsf{B}|$ in (3.33), we have

$$\frac{\tilde{L}L^2\tau^2}{\nu^2|\mathsf{B}|} \le c^2 K^{\mu-2}, \; 1 + \frac{2}{\tilde{L}\beta\tau} \le 1 + \frac{2K}{c}. \tag{3.35}$$

By the first inequality in (3.35) and the choice of β in (3.33), we have from (3.21) that

$$\Gamma_0 \leq \frac{\exp(c^2+2c)-1}{c^2+2c}K. \tag{3.36}$$

Plugging (3.35) and (3.36) into (3.20) with $k=0$, and using the choice of τ in (3.33), we have

$$\frac{\Delta_0}{\tau} \geq \nu - \frac{\nu}{2}\frac{\hat{L}}{\sqrt{\tilde{L}L}}\exp(c^2+2c)c \geq \frac{\nu}{2}, \tag{3.37}$$

where the last inequality is due to (3.34). Note that $\Delta_{\min} = \Delta_0$, we thus have

$$\Delta_{\min} \geq \nu\tau/2. \tag{3.38}$$

Second, by choosing

$$S = \left\lceil \frac{2(f(X^{0,0}) - f(X^*))}{\nu\epsilon} \cdot \frac{1}{K\tau} \right\rceil, \tag{3.39}$$

we know from (3.38) and (3.31) that $\mathbb{E}[\|\mathrm{grad}\, f(X_r)\|^2] \leq \epsilon$. From (3.33) we have $K\tau \geq \frac{c\nu\kappa^{1/3}}{\sqrt{\tilde{L}L}}n^{1/3}$, which together with (3.39) gives

$$S \leq \frac{c_1 n^{-\frac{1}{3}}}{\epsilon} + 1, \tag{3.40}$$

where $c_1 = \frac{2\sqrt{\tilde{L}L}(f(X^{0,0})-f(X^*))}{c\nu^2\kappa^{1/3}}$. Using (3.33) and noting $|\mathsf{B}| \leq 2K^{2-3\mu}$, we have

$$n + 2K|\mathsf{B}| \leq n + 2K^{2-2\mu} \leq c_2 n, \tag{3.41}$$

where $c_2 = 1 + 4\max\{1, 4\kappa^{2/3}\}$, and

$$K \leq (\kappa n)^{\frac{1}{3(1-\mu)}} + 1 \leq \big(\kappa^{\frac{1}{3(1-\mu)}} + 1\big) n^{\frac{1}{3(1-\mu)}}. \tag{3.42}$$

Thus it follows from (3.40) and (3.41) that #IFO-calls $= S(n+2K|\mathsf{B}|) = O\big(n^{2/3}/\epsilon + n\big)$. Similarly, from (3.40) and (3.42), we have #RO-calls $= 2SK = O\big(n^{\frac{\mu}{3(1-\mu)}}/\epsilon + n^{\frac{1}{3(1-\mu)}}\big)$. □

Remark 3.8. *Note that if we set the stochastic gradient $\mathcal{G}(X^k, \xi_k)$ to the exact gradient $\nabla f(X^k)$, then Algorithm 1 reduces to the deterministic gradient descent method on manifold, and the complexity becomes $O(n/\epsilon)$, which was recently established in [7].*

3.4 A S-SGD method

As a by-product of S-SVRG, we can give a vector transport-free SGD with general retraction for solving (1.2) as in Algorithm 2. The stochastic event in Line 4 is denoted by ξ_j. Note that $\mathsf{St}_{d,r}$ is compact, following the proof of Lemma 3.3, we can also show that

$$\mathbb{E}_{\xi_k}[\mathcal{G}^{\mathrm{R}}(X^k, \xi_k) - \mathrm{grad}\, f(X^k)] = 0 \text{ and } \mathbb{E}_{\xi_k}\big[\|\mathcal{G}^{\mathrm{R}}(X^k, \xi_k) - \mathrm{grad}\, f(X^k)\|_{\mathsf{F}}^2\big] \leq \sigma^2$$

for some constant σ. Using the similar techniques as in [12], we can show in Theorem 3.9 that the IFO-calls and RO-calls complexities of S-SGD are both $O(1/\epsilon^2)$. For the sake of brevity, we omit the proof here.

Theorem 3.9. *Let the step sizes $\{\tau_j\}$ be chosen as $\tau_j \equiv \min\left\{\frac{\nu}{\hat{L}}, \frac{\tilde{D}}{\sigma\sqrt{N}}\right\}$, where $\tilde{D} > 0$ is some constant. Suppose that the probability mass function $\mathbb{P}_{\bar{j}}(\cdot)$ in Algorithm 2 is chosen as $\mathbb{P}_{\bar{j}}(j) = 1/N$, $j = 0, \ldots, N-1$. It holds that*

$$\frac{1}{\hat{L}}\mathbb{E}\left[\|\mathrm{grad}\, f(X^{\bar{j}})\|_{\mathsf{F}}^2\right] \leq \frac{D_{f,\mathcal{R}}^2}{N}\left(\frac{\hat{L}}{\nu} + \frac{1}{T_{\mathcal{R}}}\right) + \left(\tilde{D} + \frac{D_{f,\mathcal{R}}^2}{\tilde{D}}\right)\frac{\sigma}{\sqrt{N}}, \tag{3.43}$$

where $D_{f,\mathcal{R}} = [2(f(X^0) - f(X^*))/\hat{L}]^{\frac{1}{2}}$ and the expectation is taken with respect to $\bar{j}$ and $\xi_{[N-1]} = (\xi_0, \xi_1, \ldots, \xi_{N-1})$. Moreover, (3.43) indicates that the IFO-calls and RO-calls complexities of Algorithm 2 for achieving a stochastic ϵ-stationary point of problem (1.2) are both $O(1/\epsilon^2)$.

Algorithm 2 A S-SGD for problem (1.2)

1: Given $X^0 \in \mathsf{St}_{d,r}$, maximal iteration number N, step sizes $\{\tau_j\}_{j\geq 0}$, and probability mass function $\mathbb{P}_{\bar{j}}(\cdot)$ supported on $\{0,\ldots,N-1\}$.
2: Generate the random variable $\bar{j}$ according to $\mathbb{P}_{\bar{j}}(\cdot)$.
3: **for** $j = 0,\ldots,\bar{j}-1$ **do**
4: Pick a random $i \in \{1,\ldots,n\}$ uniformly and compute the stochastic *Euclidean* gradient $\mathcal{G}(X^j,\xi_j)$ as $\mathcal{G}(X^j,\xi_j) = \nabla f_i(X^j)$.
5: Compute the stochastic *Riemannian* gradient as $\mathcal{G}^{\mathrm{R}}(X^j,\xi_j) = \mathbf{D}_\rho\big(X^j, \mathcal{G}(X^j,\xi_j)\big)$.
6: Update $X^{j+1} = \mathcal{R}\big(X^j, -\tau_j \mathcal{G}^{\mathrm{R}}(X^j,\xi_j)\big)$.
7: **end for**
8: **return** $X^{\bar{j}}$.

3.5 Local linear convergence of S-SVRG

We first establish the local linear convergence of S-SVRG by assuming that the Łojasiewicz inequality holds, and then we prove that it holds with high probability for low-rank matrix completion problem.

Assumption 3.10 (Łojasiewicz Inequality)**.** *For any stationary point $\bar{X} \in \mathsf{St}_{d,r}$ of problem* (1.2)*, there exist constants $\delta > 0$ and $\alpha > 0$ such that for all $\|X - \bar{X}\|_{\mathsf{F}} \leq \delta$, it holds that*

$$|f(X) - f(\bar{X})|^{1/2} \leq \alpha \|\mathrm{grad}\, f(X)\|_{\mathsf{F}}. \tag{3.44}$$

Theorem 3.11. *Assume Assumption 3.10 holds. Consider Algorithm 1 with "$X^{s+1,0} := X^{s,K}$" in Line 11 replaced by "$X^{s+1,0} := X_r^s$." Suppose that the sequence $\{X_r^s\}$ converges to a stationary point $\bar{X}$ and suppose that all the iterate points lie in the set $\{X : \|X - \bar{X}\|_{\mathsf{F}} \leq \delta\}$. We choose the parameters according to* (3.33) *and choose the probability*

$$p_{s,k} = \begin{cases} \Delta_{s,k}/(\alpha^2 + \sum_{k=0}^{K-1}\Delta_{s,k}), & k = 0,\ldots,K-1, \\ \alpha^2/(\alpha^2 + \sum_{k=0}^{K-1}\Delta_{s,k}), & k = K. \end{cases} \tag{3.45}$$

It holds that $\{X_r^s\}$ converges to $\bar{X}$ linearly in expectation, i.e.,

$$\mathbb{E}_{\xi_{s,[K-1]}}[f(X_r^s) - f(\bar{X})] \leq \frac{2\sqrt{\tilde{L}L}\alpha^2}{2\sqrt{\tilde{L}L}\alpha^2 + c\nu^2(\kappa n)^{\frac{1}{3}}}(f(X_r^{s-1}) - f(\bar{X})). \tag{3.46}$$

Proof. First, by the Łojasiewicz inequality (3.44), we obtain from (3.19) that

$$\mathbb{E}_{\xi_{s,[K-1]}}[f(X^{s,K})] \leq f(X^{s,0}) - \frac{1}{\alpha^2}\sum_{k=0}^{K-1}\Delta_k \mathbb{E}_{\xi_{[K-1]}}[f(X^{s,k}) - f(\bar{X})]. \tag{3.47}$$

Note that $\mathbb{E}_{\xi_{s,[K-1]}}[f(X_r^s) - f(\bar{X})] = \sum_{k=0}^{K} p_k \mathbb{E}_{\xi_{s,[K-1]}}[f(X^{s,k}) - f(\bar{X})]$. From (3.47) and (3.45), we have

$$\mathbb{E}_{\xi_{s,[K-1]}}[f(X_r^s) - f(\bar{X})] \leq \frac{\alpha^2}{\alpha^2 + \sum_{k=0}^{K-1}\Delta_{s,k}}(f(X^{s,0}) - f(\bar{X})). \tag{3.48}$$

From (3.33), we see that

$$\sum_{k=0}^{K-1}\Delta_{s,k} \geq \frac{K\nu\tau}{2} \geq \frac{c\nu^2}{2\sqrt{\tilde{L}L}}(\kappa n)^{\frac{1}{3}},$$

which together with (3.48) and $X^{s,0} = X_r^{s-1}$ implies (3.46). □

There are some existing works which consider the linear convergence of the SVRG on manifold. In [35], Zhang, Reddi and Sra established the linear convergence of RSVRG for geodesically convex function. However, this result is trivial because every smooth geodescially convex function on a compact Riemannian manifold is a constant (see [5]). Moreover, [35] also established the linear convergence result under the assumption that $f(X)$ is globally τ-gradient dominated, i.e., there exists $\tau > 0$ such that

$$f(X) - f(X^*) \le \tau \|\mathrm{grad}\, f(X)\|_{\mathsf{F}}^2, \quad \forall X \in \mathcal{M}, \tag{3.49}$$

where X^* is the optimal solution. However, it should be noted that (3.49) is very difficult to be verified because X^* is unknown. Kasai, Sato and Mishra [17] established the local linear convergence of R-SVRG under the assumption that the sequence converges to a non-degenerate local minimizer at which the Riemannian Hessian is positive definite. Xu and Ke [34] showed the linear convergence of their SVRRG for eigenvalue problem where they assume that the initial point is sufficiently close to the optimal solution. Note that none of the above three works gave a nontrivial example that satisfies their corresponding assumptions. Besides, Wu [32] established the local linear convergence of Stiefel-SVRG by using the fact that the Łojasiewicz inequality holds for PCA problem [20].

In the following we show that the Łojasiewicz inequality (3.44) holds locally with high probability for matrix completion problems on Grassmann manifold.

Given a rank r matrix $M \in \mathbb{R}^{m\times n}$ with $m \ge n$ and $M = U\Sigma V^{\mathsf{T}}$, where $U^{\mathsf{T}}U = mI_r$, $V^{\mathsf{T}}V = nI_r$ (note that this can be obtained by the SVD of M). The matrix completion problem aims to recover M by partial observations on a subset $\Omega \subset \{1,\ldots,m\}\times\{1,\ldots,n\}$. As is done in [18], we define

$$F(W,Z) \coloneqq \min_{S\in\mathbb{R}^{r\times r}} \mathcal{F}(W,Z,S) \quad \text{and} \quad \mathcal{F}(W,Z,S) \coloneqq \frac{1}{2}\|\mathcal{P}_\Omega(M - WSZ^{\mathsf{T}})\|_{\mathsf{F}}^2,$$

where $W \in \mathbb{R}^{m\times r}$ and $Z \in \mathbb{R}^{n\times r}$ satisfy $W^{\mathsf{T}}W = mI_r$ and $Z^{\mathsf{T}}Z = nI_r$. Consider

$$\widetilde{F}(W,Z) = F(W,Z) + \varrho\sum_{i=1}^{m} G_1\left(\frac{\|W^{(i)}\|^2}{3\mu_0 r}\right) + \varrho\sum_{j=1}^{n} G_1\left(\frac{\|Z^{(j)}\|^2}{3\mu_0 r}\right), \tag{3.50}$$

where the parameter $\varrho > 0$, μ_0 is the incoherence parameter (see, e.g., [18]) of M, $W^{(i)}$ and $Z^{(j)}$ are respectively the ith and jth columns of W^{T} and Z^{T}, and

$$G_1(z) = \begin{cases} 0, & \text{if } z \le 1, \\ e^{(z-1)^2} - 1, & \text{if } z \ge 1. \end{cases} \tag{3.51}$$

The regularized matrix completion problem is formulated as follows [18]:

$$\min \widetilde{F}(\mathbf{u}), \quad \text{s.t.} \quad \mathbf{u} \in \mathsf{M}(m,n), \tag{3.52}$$

where $\mathbf{u} = (W,Z)$ and the Cartesian product Grassmann manifold $\mathsf{M}(m,n) = \{(W \in \mathbb{R}^{m\times r}, Z \in \mathbb{R}^{n\times r})\colon W^{\mathsf{T}}W = mI_r, Z^{\mathsf{T}}Z = nI_r\}$. Note that $\widetilde{F}(\mathbf{u}) = \widetilde{F}(W,Z) \equiv \widetilde{F}(WQ_1, ZQ_2)$ for any $Q_1, Q_2 \in \mathsf{St}_r$. Define

$$\mathcal{K}(\mu') = \left\{(W,Z) : \|W^{(i)}\|^2 \le \mu' r, \|Z^{(j)}\|^2 \le \mu' r \quad \forall i \in \{1,\ldots,m\}, j \in \{1,\ldots,n\}\right\}.$$

Denote $\mathbf{u}^* = (U,V)$. If $\mathbf{u}^* \in \mathsf{M}(m,n) \cap \mathcal{K}(3\mu_0)$, from (3.51) and (3.50), we know $\widetilde{F}(u) \equiv 0$, which means that $\mathbf{u}^*$ is the optimal solution of (3.52). We now give the Łojasiewicz inequality result of (3.52), whose proof is relegated to Appendix C.

Theorem 3.12. *For $M = U\Sigma V^{\mathsf{T}}$ with $\mathbf{u}^* := (U, V) \in \mathsf{M}(m,n) \cap \mathcal{K}(3\mu_0)$. For any $\mathbf{u} \in \mathsf{M}(m,n) \cap \mathcal{K}(4\mu_0)$ and $d(\mathbf{u}, \mathbf{u}^*) \le \delta$, it holds with probability at least $1 - 1/n^4$ that*

$$\left(\widetilde{F}(\mathbf{u}) - \widetilde{F}(\mathbf{u}^*)\right)^{\frac{1}{2}} \le \left(\frac{m\Sigma_{\max}^2 + \frac{14e^{\frac{1}{9}}}{9\mu_0 r}\rho}{C\epsilon^2\Sigma_{\max}^4}\right)^{\frac{1}{2}} \|\mathrm{grad}\widetilde{F}(\mathbf{u})\|_{\mathsf{F}}, \tag{3.53}$$

where the probability is taken with respect to the uniformly random subset Ω and δ, C, ϵ are some constants. For the definition of $d(\cdot,\cdot)$ and the specific conditions on δ, C and ϵ, see equation (22) and Lemma 6.5 in [18].

4 Extensions

In this section, we extend the S-SVRG for problem (1.2) to more general manifold $\mathcal{M}$. Similar to Assumption 2.1, throughout this section, we assume that $f_i(X)$ is differentiable and $\nabla f_i(X)$ is L-Lipschitz continuous over $\mathcal{M}$. The corresponding extension of S-SGD is also possible, and we omit the details for brevity.

4.1 Some special manifolds related to $\mathsf{St}_{d,r}$

The S-SVRG method can be naturally extended to optimization with the generalized orthogonality constraints

$$\{(X_1 \in \mathbb{R}^{d_1 \times r_1}, \ldots, X_p \in \mathbb{R}^{d_p \times r_p})\colon X_i^{\mathsf{T}} M_i X_i = I_{r_i},\ i = 1, \ldots, p\} \tag{4.1}$$

where M_i is symmetric positive definite. Besides, the low-rank elliptope

$$\{X \in \mathbb{R}^{m\times m}\colon \mathrm{diag}(X) = \mathbf{1}_m,\ X = X^{\mathsf{T}} \succeq 0,\ \mathrm{rank}(X) \le r \le m\},$$

where $\mathrm{diag}(X^{\mathsf{T}}X)$ is the diagonal vector of $X^{\mathsf{T}}X$. The low-rank spectrahedron

$$\{X \in \mathbb{R}^{m\times m}\colon \mathrm{tr}(X) = 1,\ X \succeq 0,\ \mathrm{rank}(X) \le r \le m\},$$

and the oblique manifold

$$\{X \in \mathbb{R}^{m\times p}\colon \mathrm{diag}(X^{\mathsf{T}}X) = \mathbf{1}_p\},$$

where $\mathbf{1}_p \in \mathbb{R}^p$ is the all-one vector, can be seen as special cases of (4.1) by some simple transformations. Specifically, letting $X = H^{\mathsf{T}}H$ with $H \in \mathbb{R}^{r\times m}$, the low-rank elliptope and the low-rank spectrahedron can be represented as $\{H \in \mathbb{R}^{r\times m}\colon \|H_i\|_{\mathsf{F}} = 1, i = 1, \ldots, m\}$ and $\{H \in \mathbb{R}^{r\times m}\colon \|H\|_{\mathsf{F}} = 1\}$, respectively; the oblique manifold is equivalent to multiple sphere constraints as $\{X \in \mathbb{R}^{m\times p}\colon \|X_i\|_{\mathsf{F}} = 1, i = 1, \ldots, p\}$, where X_i is the ith column of X.

4.2 More general manifolds

S-SVRG (Algorithm 1) is still well-defined if $\mathsf{St}_{d,r}$ is replaced by a general Riemannian manifold $\mathcal{M}$. Since the tangent space is linear, we can still establish (3.10) and (3.11). To obtain the complexity results, we need the following assumption.

Assumption 4.1. *Consider the retraction $\mathcal{R}(t)$ with $\mathcal{R}(0) = X$ on $\mathcal{M}$. We assume that there exists some positive constants $L_1^{\mathcal{M}} \ge 1$, $L_2^{\mathcal{M}}$ and the universal positive constant $T_{\mathcal{R}}^{\mathcal{M}}$ such that*

$$\|\mathcal{R}(t) - \mathcal{R}(0)\|_{\mathsf{F}} \le L_1^{\mathcal{M}} t \|\mathcal{R}'(0)\|_{\mathsf{F}}, \tag{4.2}$$

$$\|\mathcal{R}(t) - \mathcal{R}(0) - t\mathcal{R}'(0)\|_{\mathsf{F}} \le L_2^{\mathcal{M}} t^2 \|\mathcal{R}'(0)\|_{\mathsf{F}}^2, \tag{4.3}$$

for any $t \in [0, T_{\mathcal{R}}^{\mathcal{M}}]$.

If further assuming that $\nabla f_i(X)$ is bounded on $\mathcal{M}$, then we can obtain the same complexity result as shown in Theorem 3.7. The proof is given in Appendix D.1.

Theorem 4.2. *Consider Algorithm 1 for problem* (1.1). *Suppose that Assumption 4.1 holds. Moreover, we assume that*

$$\|\nabla f_i(X)\|_{\mathsf{F}} \leq C^{\mathcal{M}}, \quad \forall X \in \mathcal{M}. \tag{4.4}$$

For any $0 \leq \mu \leq 2/3$ *and* $\kappa > 0$, *we choose*

$$K = \lceil (\kappa n)^{\frac{1}{3(1-\mu)}} \rceil,\ |\mathsf{B}| = \lceil K^{2-3\mu} \rceil,\ \beta = \frac{c}{b} K^{\mu-1},\ \tau_s \equiv bK^{-\mu}, \tag{4.5}$$

where $b = \min\left\{ \frac{c\nu}{\sqrt{\tilde{L}^{\mathcal{M}} L}}, 1, T_{\mathcal{R}}^{\mathcal{M}} \right\}$, $\tilde{L}^{\mathcal{M}} = (L_1^{\mathcal{M}})^2 + 6L_1^{\mathcal{M}} L_2^{\mathcal{M}} C^{\mathcal{M}}$ *and the constant* $c \in (0,1)$ *satisfies*

$$\frac{\hat{L}^{\mathcal{M}}}{\sqrt{\tilde{L}^{\mathcal{M}} L}} \exp(c^2 + 2c) c \leq 1. \tag{4.6}$$

Let $p_{s,k} = \frac{\Delta_{s,k}}{\sum_{k=0}^{K-1} \Delta_{s,k}}, k = 0, \ldots, K-1$ *and* $p_{s,K} = 0$, *where*

$$\frac{\Delta_{s,k}}{\tau_s} = \nu - \frac{\hat{L}^{\mathcal{M}}}{2} \tau_s \left[1 + \left(\frac{\tilde{L}_1^{\mathcal{M}} + \tilde{L}_2^{\mathcal{M}} K \tau_s}{\tilde{L}^{\mathcal{M}}} + \frac{2}{\tilde{L}^{\mathcal{M}} \beta \tau_s} \right) \frac{\tilde{L}^{\mathcal{M}} L^2 \tau_s^2}{\nu^2 |\mathsf{B}|} \Gamma_{s,k}^{\mathcal{M}} \right] \tag{4.7}$$

with

$$\Gamma_{s,k}^{\mathcal{M}} = \Gamma\Big(2\beta\tau_s + \frac{\big(\tilde{L}_1^{\mathcal{M}} + \tilde{L}_2^{\mathcal{M}} K \tau_s\big) L^2}{\nu^2 |\mathsf{B}|} \tau_s^2, K - k \Big) \tag{4.8}$$

in which $\tilde{L}_1^{\mathcal{M}} = (L_1^{\mathcal{M}})^2$ *and* $\tilde{L}_2^{\mathcal{M}} = 6L_1^{\mathcal{M}} L_2^{\mathcal{M}} C^{\mathcal{M}}$. *To obtain a stochastic* ϵ*-stationary point of* (1.2), *the IFO-calls and RO-calls complexities are* $O(n^{2/3}/\epsilon + n)$ *and* $O(n^{\frac{\mu}{3(1-\mu)}}/\epsilon + n^{\frac{1}{3(1-\mu)}})$, *respectively. In particular, when* $\mu = 0$, *i.e.,* $|\mathsf{B}| = \lceil (\kappa n)^{2/3} \rceil$, *the IFO-calls and RO-calls complexities become* $O(n^{2/3}/\epsilon + n)$ *and* $O(1/\epsilon + n^{1/3})$, *respectively.*

If the boundedness assumption (4.4) does not hold, as in [7], we need the following assumption.

Assumption 4.3 ([7])**.** *For any* $t \geq 0$, *we assume that there exits a universal positive constant* $\hat{L}^{\mathcal{M}}$ *such that*

$$f(\mathcal{R}(t)) \leq f(X) + t\langle \operatorname{grad} f(X), E \rangle_X + \frac{\hat{L}^{\mathcal{M}}}{2} t^2 \|E\|_{\mathsf{F}}^2$$

for any $t \in [0, T_{\mathcal{R}}^{\mathcal{M}}]$.

We thus can establish the iteration complexity results as follows. The proof is given in Appendix D.2.

Theorem 4.4. *Consider Algorithm 1 for problem* (1.1). *Suppose that Assumption 4.1 and Assumption 4.3 hold. For any* $0 \leq \theta \leq 1$ *and* $\kappa > 0$, *we choose*

$$K = \lceil (\kappa n)^{\frac{1}{3-2\theta}} \rceil,\ |\mathsf{B}| = \lceil K^{2-2\theta} \rceil,\ \beta = \frac{c}{K},\ \tau_s \equiv bK^{-\theta}, \tag{4.9}$$

where $b = \min\left\{ \frac{c\nu}{L_1^{\mathcal{M}} L}, 1, T_{\mathcal{R}}^{\mathcal{M}} \right\}$ *and the constant* $c \in (0,1)$ *satisfies*

$$\frac{\hat{L}^{\mathcal{M}}}{L_1^{\mathcal{M}} L} \left(\frac{c+1}{c+2} \exp(c^2 + 2c) + \frac{1}{c+2} \right) c \leq 1. \tag{4.10}$$

Let $p_{s,k} = \frac{\Delta_{s,k}}{\sum_{k=0}^{K-1}\Delta_{s,k}}, k = 0, \ldots, K-1$ *and* $p_{s,K} = 0$, *where*

$$\frac{\Delta_{s,k}}{\tau_s} = \nu - \frac{\hat{L}^{\mathcal{M}}}{2}\tau_s\left[1 + \left(1+\beta^{-1}\right)\frac{L^2(L_1^{\mathcal{M}})^2\tau_s^2}{\nu^2|\mathsf{B}|}\Gamma_{s,k}^{\mathcal{M}}\right], \tag{4.11}$$

with $\Gamma_{s,k}^{\mathcal{M}} = \Gamma\Big(\beta + \left(1+\beta^{-1}\right)\frac{L^2(L_1^{\mathcal{M}})^2\tau_s^2}{\nu^2|\mathsf{B}|}, K-k\Big)$, *To obtain a stochastic* ϵ*-stationary point of problem* (1.2)*, the IFO-calls and RO-calls complexities are* $O\big(n^{\frac{2-\theta}{3-2\theta}}/\epsilon + n\big)$ *and* $O\big(n^{\frac{\theta}{3-2\theta}}/\epsilon + n^{\frac{1}{3-2\theta}}\big)$*, respectively. In particular, when* $\theta = 0$*, i.e.,* $|\mathsf{B}| = \lceil(\kappa n)^{2/3}\rceil$*, the IFO-calls and RO-calls complexities become* $O(n^{2/3}/\epsilon + n)$ *and* $O(1/\epsilon + n^{1/3})$*, respectively.*

5 A practical S-SVRG-BB algorithm

One of the major issues in SGD is how to choose step size while running the algorithm. Recently, Tan et al. [29] proposed the SVRG-BB method, which incorporates the BB step size [4] to SVRG. The numerical results showed that SVRG-BB performs comparably to SVRG with best-tuned step sizes. BB step size was also used in optimization on Riemannian manifold (see, e.g., [13, 14, 31]). Motivated by these results, we propose to incorporate the BB step size to compute the step size τ_s in S-SVRG. Similar as [31], we define

$$\mathsf{S}^s = X^s - X^{s-1} \quad \text{and} \quad \mathsf{Y}^s = \operatorname{grad} f(X^s) - \operatorname{grad} f(X^{s-1}),$$

and compute the BB step size by

$$\tau_s^{\mathrm{LBB}} = \langle \mathsf{S}^s, \mathsf{S}^s\rangle / |\langle \mathsf{S}^s, \mathsf{Y}^s\rangle|. \tag{5.1}$$

Given $0 < \tau_{\min} < \tau_{\max}$, as done in [23], we provide safeguards for τ_s^{LBB}, namely, compute $\tau_s^{\mathrm{LBB}} = \max\{\tau_{\min}, \min\{\tau_s^{\mathrm{LBB}}, \tau_{\max}\}\}$. We set

$$\tau_s = \tau_s^{\mathrm{LBB}}/K. \tag{5.2}$$

We call Algorithm 1 with τ_s computed by (5.2) as S-SVRG-BB algorithm. By rewriting (5.2) as $\tau_s = c\nu/\sqrt{\tilde{L}}LK^{-\mu}$ with $c = \rho K^{\mu-1}$ in which

$$\rho\nu/\sqrt{\tilde{L}}L \in [\tau_{\min}, \tau_{\max}], \tag{5.3}$$

we immediately obtain the following result for S-SVRG-BB from Theorem 3.7.

Corollary 5.1. *Consider S-SVRG-BB algorithm. Given constant* $0 \le \mu \le 2/3$*, we select* K*,* $|\mathsf{B}|$*,* β *and* τ_s *by* (3.33)*, where the positive constant* κ *is chosen such that* $c = \rho K^{\mu-1}$ *(*ρ *satisfies* (5.3)*) lies in* $(0,1)$ *and satisfies* (3.34)*. To obtain a stochastic* ϵ*-stationary point of* (1.2)*, the IFO-calls and RO-calls complexities are* $O(n/\epsilon)$ *and* $O(1/\epsilon)$*, respectively.*

6 Numerical results

In this section, we compare VR-PCA [25], R-SVRG [17] and SVRRG [34] with our S-SVRG and S-SVRG-BB for solving PCA and matrix completion (MC) problems. Note that for problem (1.2), SMART-SVRG [3] is a special case of S-SVRG where the retraction is the gradient projection. If we restrict $f(X)$ to be a quadratic function (without linear term) and use the retraction of QR factorization or polar decomposition, our S-SVRG with fixed step size becomes Stiefel-SVRG [32]. We consider seven types of retractions (A.1)-(A.8) and we denote them by "qr", "pd", "wy", "jd", "gp", "exp" and "gr", respectively (see Appendix A.1

and A.2 for details). For each test instance, we run each method 20 times from random initial points. For each run, the initial random number generator seeds for different methods are the same. We stop each method at the sth epoch when $\|\mathrm{grad}\, f(X^{s,0})\|_{\mathsf{F}} \leq 10^{-6}$ or s is larger than or equal to the maximal epoch number 200. We always choose the batchsize $|\mathsf{B}| = 0.01n$ and set the maximal inner iteration number as $K = 5n$. For all the aforementioned SVRG-type methods, as done in [15], we use K iterations of the S-SGD method to improve the quality of the random initial points. For PCA problem, we set $\rho = 0$ since there always hold $X^{\mathsf{T}}\mathcal{G}(X^{s,k}, \xi_{s,k}) \equiv \mathcal{G}(X^{s,k}, \xi_{s,k})^{\mathsf{T}}X$ and $X^{\mathsf{T}}\nabla f(X) \equiv \nabla f(X)^{\mathsf{T}}X$ and thus $\mathcal{G}^{\mathrm{R}}(X^{s,k}, \xi_{s,k}) \equiv \mathbf{D}_0(X^{s,k}, \mathcal{G}(X^{s,k}, \xi_{s,k}))$ and $\mathrm{grad}\, f(X^s) \equiv \mathbf{D}_0(X^s, \nabla f(X^s))$. For MC problem, we shall choose specific ρ for different methods since $\mathcal{G}^{\mathrm{R}}(X^{s,k}, \xi_{s,k})$ depends on ρ although we always have $\mathrm{grad}\, f(X^s) \equiv \mathbf{D}_0(X^s, \nabla f(X^s))$. The parameters $\tau_{\min}$ and $\tau_{\max}$ are chosen to be 10^{-8} and 10^8, respectively. In our numerical experiments, we use a new function $\phi(t)$ in 'jd' retraction (A.4) as

$$\phi(t) = \begin{cases} t/2, & \text{if } t < 10^{-10}, \\ 1/2, & \text{if } t \geq 10^{-10}. \end{cases}$$

Note that such chosen $\phi(t)$ satisfies the condition (A.5) and it emphasis the role of $X^{\mathsf{T}}E$ in (A.4) when t is small.

Our codes were written in MATLAB (Release 2016b) and all the experiments were conducted in Ubuntu 16.04 LTS on a Dell workstation with a 3.5-GHz Intel Xeon E3-1240 v5 processor with access to 32 GB of RAM.

6.1 Principal component analysis

Given the observation data matrix $A \in \mathbb{R}^{d\times n}$, the PCA problem can be formulated as

$$\max_{X\in\mathbb{R}^{d\times r}} \frac{1}{n}\mathrm{tr}(X^{\mathsf{T}}(A-\bar{\mathbf{A}})(A-\bar{\mathbf{A}})^{\mathsf{T}}X), \quad \text{s.t.} \quad X^{\mathsf{T}}X = I_r, \tag{6.1}$$

where $\bar{\mathbf{A}} = \bar{A}\mathbf{1}_d^{\mathsf{T}}$ with $\bar{A} = \frac{1}{n}\sum_{i=1}^n A_i$ and A_i being the ith column of A. It is easy to see that (6.1) is a special case of (1.2) with $f_i(X) = -\mathrm{tr}(X^{\mathsf{T}}(A_i-\bar{A})(A_i-\bar{A})^{\mathsf{T}}X)$.

In our experiments, we generated A using to the following MATLAB code:

```
temp = [1:d].^(0.618);  A = randn(d,N);
A = bsxfun(@times,temp',A);  A = A./max(max(abs(A)));
```

We set $d = 1,000$, $n = 10,000$ and consider three choices of r: 10, 20, 40. We first compare S-SVRG with seven retractions and existing methods R-SVRG, VR-PCA and SVRRG. For these methods, we report their performance with best-tuned step sizes chosen from the set $\{1, 2, \ldots, 100\}$. The results are reported in Table 2. In this table, "retr." denotes the type of the retraction, "τ^*" denotes the best-tuned step size, "epoch" gives the minimal, average and maximal number of epoches and the standard deviation. The term "$\overline{\mathrm{nrm}}$" denotes the average F-norm of Riemannian gradient at the point returned by each method while "$\overline{\mathrm{err}}$" is the average relative function value error to the optimal value. The average CPU time "$\bar{\mathrm{t}}$", evaluated by the tic-toc commands, is in seconds. We have the following observations from Table 2. For the best-tuned step sizes, our S-SVRG is always faster than R-SVRG and SSVRG while the quality of the solution is similar. The performance of S-SVRG with the retraction of 'gp', (i.e., SMART-SVRG), is better than VR-PCA which also adopts the retraction of 'gp' but with additional twist procedures. In terms of the average number of epochs, S-SVRG with 'gp' is always the worst one among the seven tested retractions.

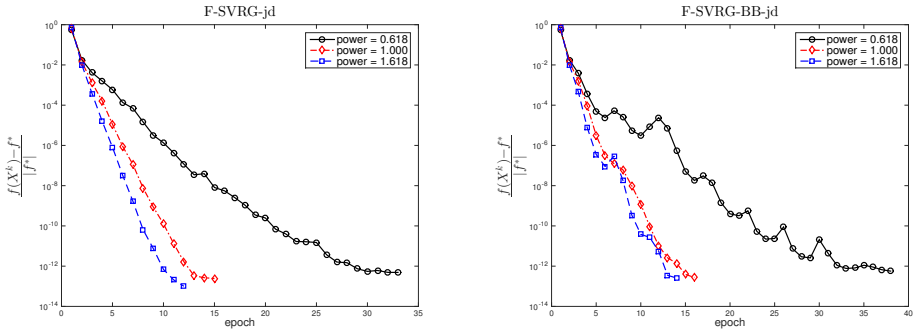
To investigate the efficiency of S-SVRG-BB, we compare the performance of S-SVRG-BB and S-SVRG with best-tuned step sizes. The comparison results are reported in Table 3. From

Table 2: Comparison of S-SVRG and existing methods for PCA instances: best-tuned step size, $d = 1000, n = 10000$.

	Existing methods						S-SVRG					
retr.	method	τ^*	epoch	$\overline{\text{nrm}}$	$\overline{\text{err}}$	$\bar{\text{t}}$	τ^*	epoch	$\overline{\text{nrm}}$	$\overline{\text{err}}$	$\bar{\text{t}}$
					$r = 10$						
exp	RSVRG	1.2	$(39, 52.9, 74, 8.2)$	9e-07	7e-11	25.2	1.2	$(41, 52.9, 72, 8.8)$	9e-07	8e-11	11.2
pd	SVRRG	1.2	$(42, 53.1, 70, 7.7)$	9e-07	7e-11	11.5	1.2	$(41, 53.0, 73, 8.8)$	9e-07	7e-11	8.7
gp	VR-PCA	1.3	$(42, 53.1, 67, 8.3)$	9e-07	7e-11	10.4	1.3	$(39, 53.4, 73, 10.1)$	9e-07	7e-11	7.8
qr	—	—	—	—	—	—	1.1	$(32, 52.9, 70, 10.9)$	9e-07	1e-10	8.2
wy	—	—	—	—	—	—	1.2	$(41, 52.8, 72, 8.8)$	9e-07	8e-11	9.8
jd	—	—	—	—	—	—	1.2	$(41, 52.8, 72, 8.8)$	9e-07	8e-11	9.9
gr	—	—	—	—	—	—	0.6	$(48, 55.0, 76, 6.8)$	9e-07	8e-11	8.1
					$r = 20$						
exp	RSVRG	1.3	$(73, 110.5, 159, 19.4)$	9e-07	5e-11	100.0	1.4	$(67, 106.2, 156, 18.5)$	9e-07	4e-11	37.3
pd	SVRRG	1.3	$(71, 113.8, 172, 22.4)$	9e-07	4e-11	35.0	1.3	$(80, 107.2, 128, 12.6)$	9e-07	5e-11	23.3
gp	VR-PCA	1.4	$(80, 109.3, 153, 18.9)$	9e-07	5e-11	30.8	1.4	$(70, 107.3, 139, 17.3)$	9e-07	5e-11	20.6
qr	—	—	—	—	—	—	1.3	$(71, 106.3, 138, 18.0)$	9e-07	5e-11	21.6
wy	—	—	—	—	—	—	1.4	$(66, 106.1, 151, 17.9)$	9e-07	4e-11	27.9
jd	—	—	—	—	—	—	1.4	$(66, 106.1, 151, 17.9)$	9e-07	5e-11	24.4
gr	—	—	—	—	—	—	0.7	$(54, 101.8, 139, 24.0)$	9e-07	5e-11	18.5
					$r = 40$						
exp	RSVRG	1.3	$(82, 104.7, 127, 13.4)$	9e-07	2e-11	182.8	1.4	$(75, 104.2, 126, 13.6)$	9e-07	2e-11	63.0
pd	SVRRG	1.3	$(85, 103.8, 131, 14.0)$	9e-07	2e-11	46.1	1.4	$(71, 101.8, 125, 16.6)$	9e-07	2e-11	34.9
gp	VR-PCA	1.4	$(73, 106.3, 164, 23.7)$	9e-07	3e-11	49.0	1.2	$(78, 111.2, 138, 18.9)$	9e-07	4e-11	35.7
qr	—	—	—	—	—	—	1.3	$(80, 98.2, 129, 14.0)$	9e-07	3e-11	32.6
wy	—	—	—	—	—	—	1.4	$(74, 103.9, 126, 13.9)$	9e-07	2e-11	46.5
jd	—	—	—	—	—	—	1.4	$(74, 103.9, 126, 13.9)$	9e-07	3e-11	38.4
gr	—	—	—	—	—	—	0.7	$(72, 103.1, 137, 18.8)$	9e-07	3e-11	27.9
					$r = 60$						
exp	RSVRG	1.2	$(78, 100.7, 148, 17.0)$	9e-07	2e-11	281.2	1.3	$(65, 88.3, 109, 12.8)$	9e-07	3e-11	82.7
pd	SVRRG	1.2	$(74, 95.4, 120, 12.6)$	9e-07	2e-11	63.9	1.3	$(61, 86.3, 109, 14.0)$	9e-07	3e-11	40.4
gp	VR-PCA	1.4	$(67, 95.0, 143, 19.4)$	9e-07	2e-11	66.3	1.3	$(70, 99.0, 130, 17.2)$	9e-07	2e-11	41.4
qr	—	—	—	—	—	—	1.4	$(77, 95.5, 118, 11.9)$	9e-07	2e-11	44.2
wy	—	—	—	—	—	—	1.3	$(65, 87.7, 109, 13.2)$	9e-07	2e-11	68.7
jd	—	—	—	—	—	—	1.3	$(65, 87.7, 109, 13.2)$	9e-07	3e-11	49.9
gr	—	—	—	—	—	—	0.7	$(64, 88.7, 113, 12.7)$	9e-07	2e-11	35.6

Table 3: Comparison of S-SVRG with best-tuned step sizes and S-SVRG-BB for PCA instances: $d = 1000, n = 10000$.

	S-SVRG					S-SVRG-BB				
retr.	τ^*	epoch	$\overline{\text{nrm}}$	$\overline{\text{err}}$	$\bar{\text{t}}$	epoch	$\overline{\text{nrm}}$	$\overline{\text{err}}$	$\bar{\text{t}}$	$\bar{\text{t}}$ratio
				$r = 10$						
exp	1.2	(41, 52.9, 72, 8.8)	9e-07	8e-11	11.2	(77, 102.7, 153, 18.0)	9e-07	4e-09	21.9	2.0
pd	1.2	(41, 53.0, 73, 8.8)	9e-07	7e-11	8.7	(77, 102.8, 153, 18.2)	9e-07	4e-09	16.9	1.9
qr	1.1	(32, 52.9, 70, 10.9)	9e-07	1e-10	8.2	(77, 103.2, 153, 18.2)	9e-07	4e-09	16.1	2.0
wy	1.2	(41, 52.8, 72, 8.8)	9e-07	8e-11	9.8	(77, 102.7, 153, 18.0)	9e-07	4e-09	19.0	1.9
jd	1.2	(41, 52.8, 72, 8.8)	9e-07	8e-11	9.9	(77, 102.7, 153, 18.0)	9e-07	4e-09	19.3	2.0
gp	1.3	(39, 53.4, 73, 10.1)	9e-07	7e-11	7.8	(77, 103.5, 153, 17.9)	9e-07	4e-09	15.0	1.9
gr	0.6	(48, 55.0, 76, 6.8)	9e-07	8e-11	8.1	(52, 75.9, 118, 15.7)	9e-07	2e-09	11.2	1.4
				$r = 20$						
exp	1.4	(67, 106.2, 156, 18.5)	9e-07	4e-11	37.3	(133, 204.5, 301, 45.6)	9e-07	5e-09	72.2	1.9
pd	1.3	(80, 107.2, 128, 12.6)	9e-07	5e-11	23.3	(133, 204.3, 300, 45.1)	1e-06	5e-09	44.2	1.9
qr	1.3	(71, 106.3, 138, 18.0)	9e-07	5e-11	21.6	(137, 207.7, 304, 47.0)	1e-06	5e-09	42.6	2.0
wy	1.4	(66, 106.1, 151, 17.9)	9e-07	4e-11	27.9	(133, 204.3, 300, 45.4)	1e-06	5e-09	53.5	1.9
jd	1.4	(66, 106.1, 151, 17.9)	9e-07	5e-11	24.4	(133, 204.3, 300, 45.4)	1e-06	5e-09	47.3	1.9
gp	1.4	(70, 107.3, 139, 17.3)	9e-07	5e-11	20.6	(138, 208.8, 307, 47.6)	9e-07	5e-09	39.8	1.9
gr	0.7	(54, 101.8, 139, 24.0)	9e-07	5e-11	18.5	(83, 146.6, 236, 39.9)	9e-07	3e-09	26.7	1.4
				$r = 40$						
exp	1.4	(75, 104.2, 126, 13.6)	9e-07	2e-11	63.0	(126, 172.4, 242, 35.6)	9e-07	2e-09	105.3	1.7
pd	1.4	(71, 101.8, 125, 16.6)	9e-07	2e-11	34.9	(126, 172.4, 242, 35.7)	9e-07	2e-09	58.9	1.7
qr	1.3	(80, 98.2, 129, 14.0)	9e-07	3e-11	32.6	(126, 172.6, 243, 35.7)	9e-07	2e-09	56.7	1.7
wy	1.4	(74, 103.9, 126, 13.9)	9e-07	2e-11	46.5	(126, 172.2, 242, 35.5)	1e-06	2e-09	77.7	1.7
jd	1.4	(74, 103.9, 126, 13.9)	9e-07	3e-11	38.4	(126, 172.2, 242, 35.5)	1e-06	3e-09	64.1	1.7
gp	1.2	(78, 111.2, 138, 18.9)	9e-07	4e-11	35.7	(127, 173.8, 268, 37.5)	9e-07	2e-09	55.9	1.6
gr	0.7	(72, 103.1, 137, 18.8)	9e-07	3e-11	27.9	(101, 140.5, 215, 30.0)	9e-07	1e-09	37.3	1.3
				$r = 60$						
exp	1.3	(65, 88.3, 109, 12.8)	9e-07	3e-11	82.7	(130, 163.9, 258, 31.7)	1e-06	2e-09	155.4	1.9
pd	1.3	(61, 86.3, 109, 14.0)	9e-07	3e-11	40.4	(130, 164.2, 255, 31.4)	1e-06	2e-09	76.5	1.9
qr	1.4	(77, 95.5, 118, 11.9)	9e-07	2e-11	44.2	(130, 164.3, 254, 31.3)	9e-07	2e-09	76.2	1.7
wy	1.3	(65, 87.7, 109, 13.2)	9e-07	2e-11	68.7	(130, 164.1, 258, 31.5)	1e-06	2e-09	129.0	1.9
jd	1.3	(65, 87.7, 109, 13.2)	9e-07	3e-11	49.9	(130, 164.1, 258, 31.5)	1e-06	2e-09	93.5	1.9
gp	1.3	(70, 99.0, 130, 17.2)	9e-07	2e-11	41.4	(133, 168.2, 272, 33.1)	9e-07	2e-09	70.5	1.7
gr	0.7	(64, 88.7, 113, 12.7)	9e-07	2e-11	35.6	(94, 130.6, 169, 24.8)	9e-07	8e-10	52.1	1.5



(a) (b)

Figure 1: The relative function value versus the number of epochs for S-SVRG and S-SVRG-BB with the retraction 'jd'

this table, we see that the S-SVRG-BB with the retraction 'jd' performs best, while the S-SVRG-BB with the retraction 'exp' performs worst, but all of them are comparable with S-SVRG with best-tuned step sizes.

Finally, we plot the relative function value $(f(X^{s,0}) - f^*)/|f^*|$ for S-SVRG and S-SVRG-BB with the retraction 'jd' in Figure 1, where f^* is the optimal function value. From this figure, we see that both S-SVRG and S-SVRG-BB converge linearly, which is consistent with the linear convergence result shown in Theorem 3.11. We have similar observation for S-SVRG and S-SVRG-BB with other retractions and we omit the figures here for brevity.

6.2 Matrix completion

Let $\Omega \in \{1, \ldots, n\} \times \{1, \ldots, n\}$. For the rank-$r$ matrix $M \in \mathbb{R}^{d\times n}$, we define the projection matrix $\mathcal{P}_\Omega(M)$ as $\mathcal{P}_\Omega(M)_{ij} = M_{ij}$ if $(i, j) \in \Omega$ and $\mathcal{P}_\Omega(M)_{ij} = 0$ otherwise. Given the observation $\mathcal{P}_\Omega(M)$, we aim to recover missing values of M by solving the following matrix completion problem [17]

$$\min_{X\in \mathsf{Gr}_{d,r}, a_i\in\mathbb{R}^r} \frac{1}{n}\sum_{i=1}^{n} \|\mathcal{P}_{\Omega_i}(Xa_i) - \mathcal{P}_{\Omega_i}(M_i)\|^2, \tag{6.2}$$

where M_i is the ith column of M, and $\mathcal{P}_{\Omega_i}(\cdot)$ is the ith column of $\mathcal{P}_\Omega(\cdot)$. It is easy to see that (6.1) is a special case of (1.2) with $f_i(X) = \min_{a_i\in\mathbb{R}^r} \|\mathcal{P}_{\Omega_i}(Xa_i) - \mathcal{P}_{\Omega_i}(M_i)\|^2$.

We generated the synthetic data matrix M as suggested in [17] with the condition number set to 10. The index set Ω is chosen randomly and uniformly from $\{1, \ldots, n\} \times \{1, \ldots, n\}$, and its size is $|\Omega| = (n + d - r)r^2$. Since the best-tuned step sizes are not easy to obtain and S-SVRG-BB has proved to be very practical, we only report the results of S-SVRG-BB. To make a fair comparison, we also adopted the BB step size (5.1) for the existing methods R-SVRG, SVRRG and VR-PCA. The corresponding methods with BB step size are named as R-SVRG-BB, SVRRG-BB and VR-PCA-BB, respectively. The method SMART-SVRG-BB stands for the SMART-SVRG with BB step size, which is essentially S-SVRG-BB with 'gp' retraction. The numerical results over 20 runs are reported in Table 4. The term '$\bar{\text{t}}$ratio' denotes the ratio of running time of each method over the minimal running time of R-SVRG-BB, SVRRG-BB, VR-PCA-BB and SMART-SVRG-BB. For instance, when $r = 20$, $d = 1000$, $n = 10000$, '$\bar{\text{t}}$ratio' for S-SVRG-BB-jd is 0.83, which means the CPU time of S-SVRG-BB-jd is only 0.83 times of that of SVRRG-BB. From Table 4 we see that using appropriate retraction, S-SVRG-BB can be faster than the four existing methods. It should be noted that since $\mathcal{R}'_{\text{gr}}(0) = -2\mathbf{D}_0(X, \nabla f(X))$ (see (A.8)), the BB step size for 'gr' is essentially enlarged by resetting $\tau_s^{\text{LBB}} := 2 \cdot \langle \mathsf{S}^s, \mathsf{S}^s\rangle / |\langle \mathsf{S}^s, \mathsf{Y}^s\rangle|$. We can also enlarge the BB step size for other retractions, and we observe that the performance is always improved. However, for sake of space, we shall not report the corresponding results.

Table 4: Comparison of R-SVRGs' for matrix completion instances: $d = 1000, n = 10000$

method	epoch	$\overline{\text{nrm}}$	$\overline{\text{err}}$	$\bar{\text{t}}$	$\bar{\text{t}}$ratio	epoch	$\overline{\text{nrm}}$	$\overline{\text{err}}$	$\bar{\text{t}}$	$\bar{\text{t}}$ratio
	$r = 10$					$r = 15$				
R-SVRG-BB	(22, 28.6, 75, 11.6)	8e-07	4e-11	222	1.21	(21, 24.4, 33, 3.9)	7e-07	5e-11	244	1.09
SVRRG-BB	(22, 24.7, 31, 2.7)	8e-07	5e-11	184	**1.00**	(20, 23.6, 29, 2.4)	7e-07	6e-11	223	**1.00**
VR-PCA-BB	(22, 26.1, 47, 5.9)	8e-07	5e-11	194	1.05	(21, 23.8, 31, 2.7)	7e-07	6e-11	227	1.02
SMART-SVRG-BB	(22, 26.6, 40, 5.7)	7e-07	5e-11	198	1.07	(21, 23.7, 31, 2.7)	7e-07	6e-11	224	1.00
S-SVRG-BB-exp	(22, 28.4, 62, 10.0)	8e-07	4e-11	214	1.16	(21, 23.9, 32, 3.0)	7e-07	6e-11	229	1.02
S-SVRG-BB-pd	(22, 25.2, 40, 4.2)	7e-07	4e-11	188	1.02	(20, 23.7, 31, 2.6)	7e-07	6e-11	224	1.00
S-SVRG-BB-qr	(22, 28.6, 48, 8.2)	7e-07	4e-11	214	1.16	(20, 24.9, 36, 4.8)	7e-07	6e-11	235	1.05
S-SVRG-BB-wy	(22, 25.7, 37, 4.5)	7e-07	4e-11	192	1.05	(20, 23.8, 31, 2.8)	7e-07	6e-11	226	1.01
S-SVRG-BB-jd	(19, 25.2, 42, 5.8)	7e-07	4e-11	188	1.02	(18, 21.4, 29, 2.6)	7e-07	6e-11	203	**0.91**
S-SVRG-BB-gr	(17, 23.5, 38, 5.6)	6e-07	8e-11	174	**0.95**	(17, 21.0, 31, 4.3)	5e-07	5e-11	197	**0.88**
	$r = 20$					$r = 25$				
R-SVRG-BB	(22, 24.3, 35, 2.9)	7e-07	5e-11	285	1.08	(22, 23.7, 27, 1.3)	7e-07	6e-11	329	1.06
SVRRG-BB	(21, 23.9, 29, 1.7)	7e-07	5e-11	265	**1.00**	(22, 23.9, 26, 1.0)	7e-07	6e-11	311	1.00
VR-PCA-BB	(22, 24.4, 35, 3.3)	7e-07	5e-11	272	1.03	(22, 23.8, 27, 1.4)	7e-07	6e-11	311	**1.00**
SMART-SVRG-BB	(22, 24.4, 35, 3.3)	7e-07	5e-11	269	1.02	(22, 23.8, 27, 1.4)	7e-07	6e-11	307	**0.99**
S-SVRG-BB-exp	(22, 24.8, 38, 4.3)	7e-07	5e-11	277	1.05	(22, 23.7, 27, 1.3)	7e-07	6e-11	313	1.01
S-SVRG-BB-pd	(21, 23.9, 29, 1.7)	7e-07	5e-11	264	**1.00**	(22, 23.9, 26, 1.0)	7e-07	6e-11	311	1.00
S-SVRG-BB-qr	(21, 24.2, 34, 2.6)	7e-07	5e-11	267	1.01	(22, 23.9, 26, 1.0)	7e-07	6e-11	310	**1.00**
S-SVRG-BB-wy	(21, 24.1, 33, 2.4)	7e-07	5e-11	268	1.01	(22, 23.9, 26, 1.0)	7e-07	6e-11	312	1.00
S-SVRG-BB-jd	(18, 19.9, 23, 1.7)	7e-07	5e-11	220	**0.83**	(18, 19.8, 25, 1.5)	6e-07	6e-11	257	**0.83**
S-SVRG-BB-gr	(17, 19.5, 29, 3.1)	5e-07	5e-11	214	**0.81**	(17, 18.1, 22, 1.6)	6e-07	7e-11	232	**0.75**

7 Conclusions

In this paper, we proposed a vector transport-free SVRG with general retraction for solving empirical risk minimization over manifold. Our S-SVRG method has several important features: (i) it can tackle general nonlinear function; (ii) it works for a variety of retractions; (iii) it formulates the unbiased and variance reduced stochastic Riemannian gradient in a simple way, without any additional costs such as parallel or vector transport. We proved that the iteration complexity of S-SVRG for obtaining a stochastic ϵ-stationary point is $O(n^{2/3}/\epsilon)$, which is far less than that of Riemannian gradient descent method. With the help of Łojasiewicz inequality, we established the linear convergence of S-SVRG. Moreover, we incorporated the BB step size to S-SVRG, and obtained a very practical S-SVRG-BB method. Numerical results on PCA and matrix completion problems showed the efficiency of the proposed methods.

Acknowledgments

We thank Bamdev Mishra, Hiroyuki Kasai and Hiroyuki Sato for sharing their codes "Riemannian_svrg".

Appendix

A Retractions on $\mathsf{St}_{d,r}$

In this section, we review several retractions on $\mathsf{St}_{d,r}$ and $\mathsf{Gr}_{d,r}$. Note that the tangent direction of the retractions of the gradient projection and gradient reflection are fixed, while other retractions have freedom to choose different directions. For a comparison of the computational cost of different retractions, see [14, 11].

A.1 Retractions on $\mathsf{St}_{d,r}$

Given $X \in \mathsf{St}_{d,r}$ and $E \in \mathbf{T}_X\mathsf{St}_{d,r}$, we next introduce six retractions along the direction E.

(i) The exponential retraction, also known as geodesic, in [10] is given as

$$\mathcal{R}_{\text{exp1}}(X, tE) = \begin{bmatrix} X & \text{qr}(D) \end{bmatrix} \exp\left(t \begin{bmatrix} X^\mathsf{T}E & -\text{upp}(D)^\mathsf{T} \\ \text{upp}(D) & 0 \end{bmatrix} \right) \begin{bmatrix} I_r \\ 0 \end{bmatrix},$$

where $D = (I_d - XX^\mathsf{T})E$ and $D = \text{qr}(D)\text{upp}(D)$ is the QR factorization of D with $\text{qr}(D) \in \mathsf{St}_{d,r}$ and $\text{upp}(D)$ being upper triangular with positive diagonal entries.

(ii) The retraction of QR factorization [1] is given as

$$\mathcal{R}_{\text{qr}}(X, tE) = \text{qr}(X + tE). \tag{A.1}$$

(iii) The retraction of polar decomposition takes the form as [21, 1]

$$\mathcal{R}_{\text{pd}}(X, tE) = \mathcal{P}_{\mathsf{St}_{d,r}}(X + tE), \tag{A.2}$$

where the projection operation $\mathcal{P}_{\mathsf{St}_{d,r}}(\tilde{X}) = \tilde{U}\tilde{V}^\mathsf{T}$, where $\tilde{U}\tilde{\Sigma}\tilde{V}^\mathsf{T}$ is the compact SVD of $\tilde{X}$. If $\tilde{X}$ is full column rank, such as when $\tilde{X} = X + tE$, $\mathcal{P}_{\mathsf{St}_{d,r}}(\tilde{X}) = \tilde{X}(\tilde{X}^\mathsf{T}\tilde{X})^{-1/2}$.

(iv) Recently, based on the Cayley transformation, Wen and Yin [31] developed a simple and efficient retraction as [4]

$$\mathcal{R}_{\text{wy}}(X, tE) = X - tU\Big(I_{2p} + \frac{t}{2}V^\mathsf{T}U\Big)^{-1}V^\mathsf{T}X, \tag{A.3}$$

where $U = [-P_XE, X]$, $V = [X,\ P_XE]$ with $P_X = I_d - \frac{1}{2}XX^\mathsf{T}$.

(v) Later on, in the point view of subspace, Jiang and Dai [14] proposed a family of retractions. For the generalized exponential retraction, generalized retraction of polar decomposition or QR factorization, see (8.2) - (8.4) in [14]. Besides, [14] also proposed a new efficient retraction as

$$\begin{cases} \mathcal{R}_{\text{jd}}(X, tE) = (2X + tD)J(t)^{-1} - X, \\ J(t) = I_r + \frac{t^2}{4}D^\mathsf{T}D - \phi(t)X^\mathsf{T}E, \end{cases} \tag{A.4}$$

where $\phi(t)$ is any function satisfying

$$\phi(0) = 0, \quad \text{and} \quad \phi'(0) = \frac{1}{2}. \tag{A.5}$$

When taking $\phi(t) = \frac{1}{2}t$, [14] showed that (A.4) and (A.3) are equivalent.

(vi) Finally, the ordinary gradient projection retraction [14, 11] is given as

$$\mathcal{R}_{\text{gp}}(X, -t\mathbf{D}_{1/4}(X, \nabla f(X))) = \mathcal{P}_{\mathsf{St}_{d,r}}(X - t\nabla f(X)). \tag{A.6}$$

Note that $\mathcal{R}'_{\text{gp}}(0) = -\mathbf{D}_{1/4}(X, \nabla f(X))$ instead of any E.

[4] It follows from Proposition 3.1 in [14] that (A.3) is well-defined if $I_r + \frac{t}{4}X^\mathsf{T}E$ is invertible. Note that this holds naturally because $X^\mathsf{T}E$ is skew-symmetric.

A.2 Retractions on $\mathsf{Gr}_{d,r}$

Given $X \in \mathsf{Gr}_{d,r}$ and $E \in \mathbf{T}_X\mathsf{Gr}_{d,r}$, the exponential retraction proposed in [1] is

$$\mathcal{R}_{\exp 2}(X, tE) = \big(X\hat{V}\cos\hat{\Sigma}t + \hat{U}\sin\hat{\Sigma}t\big)\hat{V}^{\mathsf{T}}, \tag{A.7}$$

where $E = \hat{U}\hat{\Sigma}\hat{V}^{\mathsf{T}}$ is the compact SVD of E.

Some retractions on $\mathsf{St}_{d,r}$ can be naturally taken as the retractions on $\mathsf{Gr}_{d,r}$.

Proposition A.1. *Suppose $E \neq 0$, then the retractions* (A.1) – (A.4) *can serve as the retractions on $\mathsf{Gr}_{d,r}$. If $X^{\mathsf{T}}\nabla f(X) \equiv \nabla f(X)^{\mathsf{T}}X$,* (A.6) *is also a retraction on $\mathsf{Gr}_{d,r}$.*

Proof. It only needs to show $\mathcal{R}(t) \notin [X]$ for any $t \geq 0$. We prove this by contradiction. Suppose that $Y(t_0) \in [X]$ for some $t_0 > 0$, we have $E^{\mathsf{T}}Y(t_0) = 0$.

For (A.1) and (A.2), we have $X + t_0E = \mathcal{R}_{\mathrm{qr}}(t_0)\mathrm{upp}(X + t_0E)$ and $X + t_0E = \mathcal{R}_{\mathrm{pd}}(t_0)(I_r + t_0^2E^{\mathsf{T}}E)^{\frac{1}{2}}$, respectively. For (A.4), we have $X(2I_r - J(t)) + t_0E = \mathcal{R}_{\mathrm{jd}}(t_0)J(t_0)$. By any of the above three equalities, we always have $t_0E^{\mathsf{T}}E = 0$, namely, $E = 0$. This leads to a contradiction. Thus the retractions (A.1), (A.2) and (A.4) are also well-defined retractions on $\mathsf{Gr}_{d,r}$. Note that (A.3) is equivalent to (A.4) with $\phi(t) = \frac{1}{2}t$, we immediately know that (A.3) is also a well-defined retraction on $\mathsf{Gr}_{d,r}$. If $X^{\mathsf{T}}\nabla f(X) \equiv \nabla f(X)^{\mathsf{T}}X$, the direction E for (A.6) is given as $E = -\mathbf{D}_0(X, \nabla f(X))$. With slight abuse of notation, let $U\Sigma V$ be the compact SVD of $X - t\nabla f(X)$. Then $Y(t_0) = UV^{\mathsf{T}}$ and thus $E^{\mathsf{T}}U = 0$, which further implies that $E^{\mathsf{T}}\nabla f(X) = E^{\mathsf{T}}E = 0$. This leads to a contradiction. □

Very recently, using the Householder transformation, Gao et al. [11] proposed the gradient reflection retraction as

$$\mathcal{R}_{\mathrm{gr}}(X, -2t\mathbf{D}_0(X, \nabla f(X))) = (-I_n + 2\bar{X}(\bar{X}^{\mathsf{T}}\bar{X})^{\dagger}\bar{X}^{\mathsf{T}})X, \tag{A.8}$$

where $\bar{X} = X - t\nabla f(X)$ and $(\bar{X}^{\mathsf{T}}\bar{X})^{\dagger}$ denotes the pseudo-inverse of $\bar{X}^{\mathsf{T}}\bar{X}$. By some simple computations, we can show that $\mathcal{R}'_{\mathrm{gr}}(0) = -2\mathbf{D}_0(X, \nabla f(X))$. Similar to Proposition A.1, it is easy to show that $\mathcal{R}_{\mathrm{gr}}(t)$ is a well-defined retraction on $\mathsf{Gr}_{d,r}$. The Householder transformation is also used in [27] to preserve the orthogonality constraints.

A.3 Estimation of L_1 and L_2 for polar decomposition

Lemma A.2. *For any $X \in \mathsf{St}_{d,r}$ and $E \in \mathbf{T}_X\mathsf{St}_{d,r}$, consider the retraction of polar decomposition* (A.2). *Then equations* (3.1) *and* (3.2) *hold for any $t \geq 0$ with $L_1 = 1$ and $L_2 = 1/2$.*

Proof. First we naturally have $\mathcal{R}'(0) = E$. For simplicity, denote $H = \big(I_r + t^2E^{\mathsf{T}}E\big)^{\frac{1}{2}}$. Thus we have $\mathcal{R}(t) - \mathcal{R}(0) = \big(X(I_r - H) + tE\big)H^{-1}$, which together with the fact that $\mathrm{tr}(X^{\mathsf{T}}ES) = \mathrm{tr}(SE^{\mathsf{T}}X) = 0$ for any symmetric $S \in \mathbb{R}^{r\times r}$ implies that

$$\|\mathcal{R}(t) - \mathcal{R}(0)\|_{\mathsf{F}}^2 = 2\mathrm{tr}(I_r - H^{-1}) = 2\sum_{i=1}^{r}\Big(1 - (1 + t^2\lambda_i(E^{\mathsf{T}}E))^{-1/2}\Big) \leq t^2\|E\|_{\mathsf{F}}^2, \tag{A.9}$$

where the first equality is due to $t^2E^{\mathsf{T}}E = H^2 - I_r$ which follows from the definition of H, and the inequality is due to $2(1 - (1+z)^{-1/2}) \leq z$ with $z = t^2\lambda_i(E^{\mathsf{T}}E)$.

Note that $\mathcal{R}(t) - \mathcal{R}(0) - t\mathcal{R}'(0) = (X + tE)(H^{-1} - I_r)$. Again from $t^2E^{\mathsf{T}}E = H^2 - I_r$, we have

$$\begin{aligned}\|\mathcal{R}(t) - \mathcal{R}(0) - t\mathcal{R}'(0)\|_{\mathsf{F}}^2 = \mathrm{tr}((I_r - H)^2) &= \sum_{i=1}^{r}\Big(1 - \sqrt{1 + t^2\lambda_i(E^{\mathsf{T}}E)}\Big)^2 \\ &\leq \frac{t^4}{4}\sum_{i=1}^{r}\lambda_i^2(E^{\mathsf{T}}E) \leq \frac{t^4}{4}\|E\|_{\mathsf{F}}^4,\end{aligned} \tag{A.10}$$

where the first inequality is due to $(1 - (1+z)^{1/2})^2 \leq z^2/4$ with $z = t^2\lambda_i(E^{\mathsf{T}}E)$. It follows from Lemma A.9 and Lemma A.10 that (3.1) and (3.2) hold with $L_1 = 1$ and $L_2 = \frac{1}{2}$, respectively. □

A.4 Estimation of L_1 and L_2 for QR factorization

For any $A \in \mathbb{R}^{n\times n}$, as the same in [8], we define the upper triangular matrix $\mathrm{up}(A) \in \mathbb{R}^{n\times n}$ as $\mathrm{up}(A)_{ij} = A_{ij}$ if $i < j$, $\mathrm{up}(A)_{ij} = A_{ii}/2$ if $i = j$ and $\mathrm{up}(A)_{ij} = 0$ if $i > j$. We further have that $2\|\mathrm{up}(A)\|_{\mathsf{F}}^2 = \|A\|_{\mathsf{F}}^2 - \frac{1}{2}\sum_{i=1}^n A_{ii}^2 \le \|A\|_{\mathsf{F}}^2$, which implies that $\|\mathrm{up}(A)\|_{\mathsf{F}} \le \sqrt{2}/2\|A\|_{\mathsf{F}}$.

Lemma A.3. *For any $X \in \mathsf{St}_{d,r}$ and $E \in \mathbf{T}_X\mathsf{St}_{d,r}$, consider the retraction of QR factorization* (A.1)*. Then equations* (3.1) *and* (3.2) *hold for any $t \ge 0$ with $L_1 = 1+\sqrt{2}/2$ and $L_2 = \sqrt{10}/2$.*

Proof. Let the QR factorization of $X + tE$ be

$$X + tE = Q(t)R(t), \tag{A.11}$$

where $Q(t) \in \mathsf{St}_{d,r}$ and $R(t) \in \mathbb{R}^{r\times r}$ is upper triangular with positive diagonal elements. We then have $\mathcal{R}(t) = Q(t)$ and

$$R(t)^{\mathsf{T}}R(t) = I_r + t^2E^{\mathsf{T}}E. \tag{A.12}$$

Differentiating both sides of (A.12) with respect to t, we have $R'(t)^{\mathsf{T}}R(t)+R(t)^{\mathsf{T}}R'(t) = 2tE^{\mathsf{T}}E$ and further $(R'(t)R(t)^{-1})^{\mathsf{T}} + R'(t)R(t)^{-1} = 2tR(t)^{-\mathsf{T}}E^{\mathsf{T}}ER(t)^{-1}$. Noting that $R'(t)R(t)^{-1}$ is upper triangular, so we obtain

$$R'(t) = 2t\,\mathrm{up}\left(R(t)^{-\mathsf{T}}E^{\mathsf{T}}ER(t)^{-1}\right)R(t). \tag{A.13}$$

Differentiating both sides of (A.11) with respect to t, it follows from (A.13) that

$$Q'(t) = ER(t)^{-1} - 2t\,Q(t)\mathrm{up}\left(R(t)^{-\mathsf{T}}E^{\mathsf{T}}ER(t)^{-1}\right). \tag{A.14}$$

We now bound the term $t\|\mathrm{up}\left(R(t)^{-\mathsf{T}}E^{\mathsf{T}}ER(t)^{-1}\right)\|_{\mathsf{F}}$. Using (A.12), it is easy to verify

$$\begin{aligned} t^2\|R(t)^{-\mathsf{T}}E^{\mathsf{T}}ER(t)^{-1}\|_{\mathsf{F}}^2 &= \sum_{i=1}^r\left(\frac{t\lambda_i(E^{\mathsf{T}}E)}{1+t^2\lambda_i(E^{\mathsf{T}}E)}\right)^2 \le \sum_{i=1}^r \min\left\{t^2\lambda_i^2(E^{\mathsf{T}}E), \lambda_i(E^{\mathsf{T}}E)/4\right\} \\ &\le \|E\|_{\mathsf{F}}^2\min\left\{t^2\|E\|_{\mathsf{F}}^2, 1/4\right\}, \end{aligned} \tag{A.15}$$

where the first inequality uses $1 + t^2\lambda_i(E^{\mathsf{T}}E) \ge 2t\sqrt{\lambda_i(E^{\mathsf{T}}E)}$. Squaring both sides of (A.15) and using $\|\mathrm{up}(A)\|_{\mathsf{F}} \le \sqrt{2}/2\|A\|_{\mathsf{F}}$, we obtain

$$t\|\mathrm{up}\Big(R(t)^{-\mathsf{T}}E^{\mathsf{T}}ER(t)^{-1}\Big)\|_{\mathsf{F}} \le (\sqrt{2}/2)\|E\|_{\mathsf{F}}\min\left\{t\|E\|_{\mathsf{F}}, 1/2\right\}, \tag{A.16}$$

which together with (A.14) and (A.13), respectively, indicates

$$\|Q'(t)\|_{\mathsf{F}} \le (1+\sqrt{2}/2)\|E\|_{\mathsf{F}} \tag{A.17}$$

and

$$\|R'(t)\|_{\mathsf{F}} \le \sqrt{2}\|E\|_{\mathsf{F}}\min\left\{t\|E\|_{\mathsf{F}}, 1/2\right\}\sqrt{1+t^2\|E\|_{\mathsf{F}}^2} \le (\sqrt{10}/2)t\|E\|_{\mathsf{F}}^2. \tag{A.18}$$

By the Mean-Value Theorem, there exists $u \in (0,t)$ such that $\mathcal{R}(t) - \mathcal{R}(0) = Q(t) - Q(0) = tQ'(u)$. Then $\|\mathcal{R}(t) - \mathcal{R}(0)\|_{\mathsf{F}} = t\|Q'(u)\|_{\mathsf{F}}$, which together with (A.17) implies that (3.1) holds with $L_1 = 1+\sqrt{2}/2$. Again by the Mean-Value Theorem, noting that (A.11) and $R(0) = I_r$, we have $\mathcal{R}(t) - \mathcal{R}(0) - tE = Q(t)(R(0) - R(t)) = tQ(t)R'(u)$, where $u \in (0,t)$. Then $\|\mathcal{R}(t) - \mathcal{R}(0) - tE\|_{\mathsf{F}} \le t\|R'(u)\|_{\mathsf{F}}$, which with (A.18) yields that (3.1) holds with $L_2 = \sqrt{10}/2$. □

B Proof of Lemma 3.4

Proof. By recursively using (3.17) and noting that $\mathbf{b}_0 = 0$, we have

$$\mathbf{b}_{k+1} \leq \mathbf{a} \sum_{i=0}^{k} \mathbf{b}^{k-i} \mathbf{a}_i. \tag{B.1}$$

holds for any $k = 0, \ldots, K-1$. Again note that $\mathbf{b}_0 = 0$ we thus have from (B.1) that

$$\sum_{k=0}^{K-1} \mathbf{b}_k = \sum_{k=0}^{K-2} \mathbf{b}_{k+1} \leq \mathbf{a} \sum_{k=0}^{K-2} \sum_{i=0}^{k} \mathbf{b}^{k-i} \mathbf{a}_i = \mathbf{a} \sum_{k=0}^{K-2} \left(\sum_{i=0}^{K-2-k} \mathbf{b}^i \right) \mathbf{a}_k = \mathbf{a} \sum_{k=0}^{K-1} \frac{\mathbf{b}^{K-1-k} - 1}{\mathbf{b} - 1} \mathbf{a}_k, \tag{B.2}$$

where the last inequality is due to $\mathbf{a}_k \geq 0$. Recursively applying (3.16) yields

$$\mathbf{f}_K \leq \mathbf{f}_0 - \mathbf{c} \sum_{k=0}^{K-1} \mathbf{a}_k + \mathbf{d} \sum_{k=0}^{K-1} \mathbf{b}_k, \tag{B.3}$$

which together with (B.2) yields (3.18). □

C Proof of Theorem 3.12

First, define $d_c(\mathbf{u}, \mathbf{u}^*) = (d_c(W, U)^2 + d_c(Z, V)^2)^{\frac{1}{2}}$ with

$$d_c(W, U) = \frac{1}{\sqrt{n}} \min_{Q_1, Q_2 \in \mathsf{St}_r} \|WQ_1 - UQ_2\|_{\mathsf{F}} = \frac{1}{\sqrt{n}} \min_{Q \in \mathsf{St}_r} \|WQ - U\|_{\mathsf{F}}. \tag{C.1}$$

It is known that $d_c(\mathbf{u}, \mathbf{u}^*) \leq d(\mathbf{u}, \mathbf{u}^*)$ (see, e.g., Remark 6.1 in [18]). We now present a useful proposition.

Proposition C.1. *Suppose that* $\mathbf{u}^* \in \mathsf{M}(m, n) \cap \mathcal{K}(3\mu_0)$. *Then*

$$\widetilde{F}(\mathbf{u}) \equiv \widetilde{F}(W, Z) \leq n \left(m\Sigma_{\max}^2 + \frac{14 e^{\frac{1}{9}}}{9\mu_0 r} \varrho \right) d(\mathbf{u}, \mathbf{u}^*)^2 \tag{C.2}$$

holds for all $\mathbf{u} \in \mathsf{M}(m, n) \cap \mathcal{K}(4\mu_0)$ *with* $\mathsf{M}(m, n) = \mathsf{g}(m, r) \times \mathsf{g}(n, r)$.

Proof. First, we have

$$\begin{aligned} F(W, Z) &= \frac{1}{2} \|\mathcal{P}_\Omega(M - WSZ^{\mathsf{T}})\|_{\mathsf{F}}^2 \leq \frac{1}{2} \|\mathcal{P}_\Omega(M - W\Sigma Z^{\mathsf{T}})\|_{\mathsf{F}}^2 \\ &\leq \frac{1}{2} \|M - W\Sigma Z^{\mathsf{T}}\|_{\mathsf{F}}^2 = \frac{1}{2} \|U\Sigma V^{\mathsf{T}} - W\Sigma V^{\mathsf{T}} + W\Sigma V^{\mathsf{T}} - W\Sigma Z^{\mathsf{T}}\|_{\mathsf{F}}^2 \\ &\leq \|(U - W)\Sigma V^{\mathsf{T}}\|_{\mathsf{F}}^2 + \|W\Sigma(V - Z)^{\mathsf{T}}\|_{\mathsf{F}}^2 \\ &\leq m\Sigma_{\max}^2 \left(\|U - W\|_{\mathsf{F}}^2 + \|V - Z\|_{\mathsf{F}}^2 \right), \end{aligned} \tag{C.3}$$

where the first inequality is due to the optimality of S.

Now, let us bound the last two terms in (3.50). It is easy to show that

$$G_1(z) \leq e^{\frac{1}{9}} (z - 1)^2, \quad \forall z \in [0, 4/3]. \tag{C.4}$$

Note that $\mathbf{u} \in \mathsf{M}(m, n) \cap \mathcal{K}(4\mu_0)$ implies that $\frac{\|W^{(i)}\|^2}{3\mu_0 r} \leq \frac{4}{3}$. Define

$$\mathcal{I}_1 = \left\{ i : \frac{\|W^{(i)}\|^2}{3\mu_0 r} \leq 1, i \in \{1, \ldots, m\} \right\}, \ \mathcal{I}_2 = \left\{ i : 1 < \frac{\|W^{(i)}\|^2}{3\mu_0 r} \leq \frac{4}{3}, i \in \{1, \ldots, m\} \right\}.$$

It follows from (3.51) and (C.4) that

$$\begin{aligned}\sum_{i=1}^{m} G_1\left(\frac{\|W^{(i)}\|^2}{3\mu_0 r}\right) &= \sum_{i\in\mathcal{I}_2} G_1\left(\frac{\|W^{(i)}\|^2}{3\mu_0 r}\right)\\ &\le e^{\frac{1}{9}}\sum_{i\in\mathcal{I}_2}\left(\frac{\|W^{(i)}\|^2}{3\mu_0 r}-1\right)^2 \le e^{\frac{1}{9}}\sum_{i\in\mathcal{I}_2}\left(\frac{\|W^{(i)}\|^2}{3\mu_0 r}-\frac{\|U^{(i)}\|^2}{3\mu_0 r}\right)^2\\ &= \frac{e^{\frac{1}{9}}}{9\mu_0^2 r^2}\sum_{i\in\mathcal{I}_2}\Big(\|W^{(i)}\|-\|U^{(i)}\|\Big)^2\Big(\|W^{(i)}\|+\|U^{(i)}\|\Big)^2\\ &\le \frac{2e^{\frac{1}{9}}}{9\mu_0^2 r^2}\sum_{i\in\mathcal{I}_2}\Big(\|W^{(i)}-U^{(i)}\|^2\Big)\Big(\|W^{(i)}\|^2+\|U^{(i)}\|^2\Big)\\ &\le \frac{14e^{\frac{1}{9}}}{9\mu_0 r}\sum_{i\in\mathcal{I}_2}\Big(\|W^{(i)}-U^{(i)}\|^2\Big) \le \frac{14e^{\frac{1}{9}}}{9\mu_0 r}\|W-U\|_{\mathsf{F}}^2, \end{aligned} \tag{C.5}$$

where the second inequality is due to $\frac{\|U^{(i)}\|^2}{3\mu_0 r}\le 1$, and the fourth inequality uses the facts that $\frac{\|U^{(i)}\|^2}{3\mu_0 r}\le 1$ and $\frac{\|W^{(i)}\|^2}{3\mu_0 r}\le \frac{4}{3}$. Similarly, we have

$$\sum_{j=1}^{n} G_1\left(\frac{\|Z^{(j)}\|^2}{3\mu_0 r}\right) \le \frac{14e^{\frac{1}{9}}}{9\mu_0 r}\|Z-V\|_{\mathsf{F}}^2. \tag{C.6}$$

Combining (C.3), (C.5) and (C.6), we have

$$\widetilde{F}(\mathbf{u}) \equiv \widetilde{F}(W,Z) \le \left(m\Sigma_{\max}^2 + \frac{14e^{\frac{1}{9}}}{9\mu_0 r}\varrho\right)\left(\|U-W\|_{\mathsf{F}}^2+\|V-Z\|_{\mathsf{F}}^2\right)$$

for any $\mathbf{u}\in\mathsf{M}(m,n)\cap\mathcal{K}(4\mu_0)$. Note that $\widetilde{F}(W,Z)\equiv\widetilde{F}(WQ_W,ZQ_Z)$ for any $Q_W,Q_Z\in\mathsf{St}_r$. By the definition (C.1) of $d_c(\mathbf{u},\mathbf{u}^*)$ and $d_c(\mathbf{u},\mathbf{u}^*)\le d(\mathbf{u},\mathbf{u}^*)$, we arrive at (C.2). □

Second, it follows from Lemma 6.5 in [18] that

$$\|\mathrm{grad}\widetilde{F}(\mathbf{u})\|^2 \ge Cn\epsilon^2\Sigma_{\min}^4 d(\mathbf{u},\mathbf{u}^*)^2 \tag{C.7}$$

for all $\mathbf{u}\in\mathsf{M}(m,n)\cap\mathcal{K}(4\mu_0)$ and $d(\mathbf{u},\mathbf{u}^*)\le\delta$ with probability at least $1-1/n^4$.

Finally, combing (C.7) and (C.2), noting that $\widetilde{F}(\mathbf{u}^*)=0$, we have Theorem 3.12.

D Proofs for Theorem 4.2 and Theorem 4.4

D.1 Proof for Theorem 4.2

For fixed s, we again drop the subscript s for simplicity. Similar to Lemma 3.5, we have

$$\mathbb{E}_{\xi_{[K-1]}}[f(X^K)] \le f(X^0) - \sum_{k=0}^{K-1}\Delta_k\mathbb{E}_{\xi_{[K-1]}}\big[\|\mathrm{grad}\, f(X^k)\|_{\mathsf{F}}^2\big], \tag{D.1}$$

where Δ_k is given in (4.7). The proof of (D.1) is the same as that of Lemma 3.5 except that $\|X^k-X^0\|_{\mathsf{F}}\le 2\sqrt{r}$ is replaced by $\|X^k-X^0\|_{\mathsf{F}}\le 3C^{\mathcal{M}}L_1^{\mathcal{M}}K$, because

$$\begin{aligned}\|X^k-X^0\|_{\mathsf{F}} &\le \sum_{j=1}^{k}\|X^j-X^{j-1}\|_{\mathsf{F}} = \sum_{j=1}^{k}\|\mathcal{R}\big(X^{j-1},-\tau\mathcal{G}^{\mathrm{R}}(X^{j-1},\xi_{j-1})\big)-\mathcal{R}(X^{j-1},0)\|_{\mathsf{F}}\\ &\le \sum_{j=1}^{k}L_1^{\mathcal{M}}\tau\|\mathcal{G}^{\mathrm{R}}(X^{j-1},\xi_{j-1})\|_{\mathsf{F}} \le \sum_{j=1}^{k}L_1^{\mathcal{M}}\tau\|\mathcal{G}(X^{j-1},\xi_{j-1})\|_{\mathsf{F}} \le 3C^{\mathcal{M}}L_1^{\mathcal{M}}K\tau,\end{aligned}$$

where the second inequality is due to (4.2).

Note that Theorem 3.6 still holds. Next we estimate $\Delta_{\min}$. Again we have $\Delta_{\min} = \Delta_0$. Note that $\tilde{L}^{\mathcal{M}} = \tilde{L}_1^{\mathcal{M}} + \tilde{L}_2^{\mathcal{M}}$ and $\tilde{L}_1^{\mathcal{M}} \geq 1$, together with (4.5) and $0 \leq \mu \leq 2/3$, we obtain

$$\frac{\tilde{L}^{\mathcal{M}} L^2 \tau^2}{\nu^2 |\mathsf{B}|} \leq c^2 K^{\mu-2}, \quad \frac{\tilde{L}_1^{\mathcal{M}} + \tilde{L}_2^{\mathcal{M}} K\tau}{\tilde{L}^{\mathcal{M}}} + \frac{2}{\tilde{L}^{\mathcal{M}} \beta \tau} \leq K^{1-\mu} + \frac{2K}{c}. \tag{D.2}$$

With the first assertion in (D.2) and (4.5), we have

$$\Gamma_0^{\mathcal{M}} \leq \frac{\exp(c^2+2c)-1}{c^2+2c} K. \tag{D.3}$$

Using (D.2) and (D.3), by direct calculations, we obtain from (4.8) with $k=0$ and (4.5) that

$$\frac{\Delta_0}{\tau} \geq \nu - \frac{\nu}{2} \frac{\hat{L}^{\mathcal{M}}}{\sqrt{\tilde{L}^{\mathcal{M}}} L} \exp(c^2+2c) c \geq \frac{1}{2}\nu,$$

where the second inequality is due to (4.6).

Similar to the proof for Theorem 3.7, we arrive at Theorem 4.2.

D.2 Proof for Theorem 4.4

For fixed s, we again drop the subscript s for simplicity. Similar to Lemma 3.5, we have

$$\mathbb{E}_{\xi_{[K-1]}}[f(X^K)] \leq f(X^0) - \sum_{k=0}^{K-1} \Delta_k \mathbb{E}_{\xi_{[K-1]}} \big[\|\mathrm{grad}\, f(X^k)\|_{\mathsf{F}}^2 \big], \tag{D.4}$$

where Δ_k is given in (4.11). The proof of (D.4) is the same as that of Lemma 3.5 except that (3.27) is replaced by

$$\begin{aligned} \|X^{k+1} - X^0\|_{\mathsf{F}}^2 \leq\ & (1+\beta) \|X^k - X^0\|_{\mathsf{F}}^2 + \Big(1 + \frac{1}{\beta}\Big) \|\mathcal{R}\big(X^k, -\tau \mathcal{G}^{\mathrm{R}}(X^k, \xi_k)\big) - \mathcal{R}\big(X^k, 0\big)\|_{\mathsf{F}}^2 \\ \leq\ & (1+\beta) \|X^k - X^0\|_{\mathsf{F}}^2 + (L_1^{\mathcal{M}} \tau)^2 \Big(1 + \frac{1}{\beta}\Big) \|\mathcal{G}^{\mathrm{R}}(X^k, \xi_k)\|_{\mathsf{F}}^2, \end{aligned} \tag{D.5}$$

where the first inequality is due to the Cauchy-Schwarz inequality, $X^k = \mathcal{R}(X^k, 0)$ and (3.9), and the second inequality is due to (4.2).

Note that Theorem 3.6 still holds. Next we estimate $\Delta_{\min}$. Again we have $\Delta_{\min} = \Delta_0$. We can obtain from (4.9) that

$$\frac{L^2 (L_1^{\mathcal{M}})^2 \tau^2}{\nu^2 |\mathsf{B}|} \leq \frac{c^2}{K^2}, \ \Gamma_0^k \leq \frac{\exp(c^2+2c)-1}{c+2} K. \tag{D.6}$$

Noting that $\tau \leq c\nu/(L L_1^{\mathcal{M}})$, by some simple calculations, we know from (D.6), (4.9), (4.10) and (4.11) with $k=0$ that

$$\frac{\Delta_0}{\tau} \geq \nu - \frac{\nu}{2} \frac{\hat{L}^{\mathcal{M}}}{L L_1^{\mathcal{M}}} \left(\frac{c+1}{c+2} \exp(c^2+2c) + \frac{1}{c+2} \right) c \geq \frac{\nu}{2}, \tag{D.7}$$

and thus $\Delta_{\min} \geq \nu\tau/2$.

Similar to the proof for Theorem 3.7, we arrive at Theorem 4.4.

References

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

[2] Z. Allen-Zhu and E. Hazan, *Variance reduction for faster non-convex optimization*, arXiv:1603.05643, (2016).

[3] A. Aravkin and D. Davis, *A SMART stochastic algorithm for nonconvex optimization with applications to robust machine learning*, arXiv:1610.01101, (2016).

[4] J. Barzilai and J. M. Borwein, *Two-point step size gradient methods*, IMA J. Numer. Anal., 8 (1988), pp. 141–148.

[5] R. L. Bishop and B. O'Neill, *Manifolds of negative curvature*, T. Am. Math. Soc., 145 (1969), pp. 1–49.

[6] N. Boumal and P.-A. Absil, *Low-rank matrix completion via preconditioned optimization on the Grassmann manifold*, Linear Algebra Appl., 475 (2015), pp. 200–239.

[7] N. Boumal, P.-A. Absil, and C. Cartis, *Global rates of convergence for nonconvex optimization on manifolds*, arXiv:1605.08101, (2016).

[8] X.-W. Chang, C. C. Paige, and G. W. Stewart, *Perturbation analyses for the QR factorization*, SIAM Journal on Matrix Analysis and Applications, 18 (1997), pp. 775–791.

[9] X. Dai, Z. Liu, and A. Zhou, *A conjugate gradient optimization method for electronic structure calculations*, arXiv:1601.07676, (2016).

[10] A. Edelman, T. A. Arias, and S. T. Smith, *The geometry of algorithms with orthogonality constraints*, SIAM J. Matrix Anal. Appl., 20 (1998), pp. 303–353.

[11] B. GAO, X. LIU, X. CHEN, and Y. YUAN, *A new first-order framework for orthogonal constrained optimization problems*, Optimization Online preprint, (2016), pp. 09–5660.

[12] S. Ghadimi and G. Lan, *Stochastic first-and zeroth-order methods for nonconvex stochastic programming*, SIAM J. Optim., 23 (2013), pp. 2341–2368.

[13] B. IANNAZZO and M. PORCELLI, *The Riemannian Barzilai-Borwein method with nonmonotone line-search and the matrix geometric mean computation*, Preprint at Optimization Online, (2015).

[14] B. Jiang and Y.-H. Dai, *A framework of constraint preserving update schemes for optimization on Stiefel manifold*, Math. Program., 153 (2015), pp. 535–575.

[15] R. Johnson and T. Zhang, *Accelerating stochastic gradient descent using predictive variance reduction*, in Adv. Neural Inf. Process. Syst. 26, NIPS'13, USA, 2013, Curran Associates Inc., pp. 315–323.

[16] I. Jolliffe, *Principal Component Analysis*, John Wiley & Sons, Ltd, 2014.

[17] H. Kasai, H. Sato, and B. Mishra, *Riemannian stochastic variance reduced gradient on Grassmann manifold*, arXiv:1605.07367, (2016).

[18] R. H. Keshavan, A. Montanari, and S. Oh, *Matrix completion from a few entries*, IEEE Trans. Inf. Theory, 56 (2010), pp. 2980–2998.

[19] J. Konečnỳ, J. Liu, P. Richtárik, and M. Takáč, *Mini-batch semi-stochastic gradient descent in the proximal setting*, IEEE J. Sel. Top. Signa., 10 (2016), pp. 242–255.

[20] H. Liu, W. Wu, and A. M.-C. So, *Quadratic optimization with orthogonality constraints: Explicit Łojasiewicz exponent and linear convergence of line-search methods*, in Proc. 33rd Int. Conf. on Mach. Learn., ICML'16, JMLR.org, 2016, pp. 1158–1167.

[21] J. H. Manton, *Optimization algorithms exploiting unitary constraints*, IEEE Trans. Signal Process, 50 (2002), pp. 635–650.

[22] C. Pölitz, W. Duivesteijn, and K. Morik, *Interpretable domain adaptation via optimization over the Stiefel manifold*, Mach. Learn., 104 (2016), pp. 315–336.

[23] M. Raydan, *The Barzilai and Borwein gradient method for the large scale unconstrained minimization problem*, SIAM J. Optim., 7 (1997), pp. 26–33.

[24] S. J. Reddi, A. Hefny, S. Sra, B. Póczós, and A. Smola, *Stochastic variance reduction for nonconvex optimization*, arXiv:1603.06160, (2016).

[25] O. Shamir, *Fast stochastic algorithms for SVD and PCA: Convergence properties and convexity*, arXiv:1507.08788, (2015).

[26] ——, *A stochastic PCA and SVD algorithm with an exponential convergence rate*, in Proc. 32nd Int. Conf. on Mach. Learn., ICML'15, JMLR.org, 2015, pp. 144–152.

[27] C. Sun, Y. Yang, and Y. Yuan, *Low complexity interference alignment algorithms for desired signal power maximization problem of mimo channels*, EURASIP J. Adv. Signal Process., 2012 (2012), pp. 1–13.

[28] J. Sun, Q. Qu, and J. Wright, *Complete dictionary recovery over the sphere II: recovery by Riemannian trust-region method*, CoRR, abs/1511.04777 (2015).

[29] C. Tan, S. Ma, Y.-H. Dai, and Y. Qian, *Barzilai-Borwein step size for stochastic gradient descent*, in Adv. Neural Inf. Process. Syst. 29, NIPS'16, Curran Associates, Inc., 2016, pp. 685–693.

[30] F. J. Theis, T. P. Cason, and P. A. Absil, *Soft dimension reduction for ICA by joint diagonalization on the Stiefel manifold*, in Independent Component Analysis and Signal Separation: 8th International Conference, ICA 2009, Paraty, Brazil, March 15-18, 2009. Proceedings, T. Adali, C. Jutten, J. M. T. Romano, and A. K. Barros, eds., Springer Berlin Heidelberg, Berlin, Heidelberg, 2009, pp. 354–361.

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

[32] W. Wu, *Quadratic Optimization with Orthogonality Constraints: Explicit Łojasiewicz Exponent and Linear Convergence*, PhD thesis, The Chinese University of Hong Kong, 2016.

[33] L. Xiao and T. Zhang, *A proximal stochastic gradient method with progressive variance reduction*, SIAM J. Optim., 24 (2014), pp. 2057–2075.

[34] Z. Xu and Y. Ke, *Stochastic variance reduced Riemannian eigensolver*, arXiv:1605.08233, (2016).

[35] H. Zhang, S. J. Reddi, and S. Sra, *Fast stochastic optimization on Riemannian manifolds*, arXiv:1605.07147, (2016).