A Fast Stochastic Riemannian Eigensolver

Zhiqiang Xu
King Abdullah University of Science and Technology (KAUST)
CEMSE Division
Thuwal, 23955, Saudi Arabia

Yiping Ke
Nanyang Technological University
Singapore

Xin Gao
King Abdullah University of Science and Technology (KAUST)
CEMSE Division
Thuwal, 23955, Saudi Arabia

Abstract

We propose a fast stochastic Riemannian gradient eigensolver for a real and symmetric matrix, and prove its local, eigengap-dependent and linear convergence. The fast convergence is brought by deploying the variance reduction technique which was originally developed for the Euclidean strongly convex problems. In this paper, this technique is generalized to Riemannian manifolds for solving the geodesically non-convex problem of finding a group of top eigenvectors of such a matrix. We first propose the general variance reduction form of the stochastic Riemannian gradient, giving rise to the stochastic variance reduced Riemannian gradient method (SVRRG). It turns out that the operation of vector transport is necessary in addition to using Riemannian gradients and retraction operations. We then specialize it to the problem in question resulting in our SVRRG-EIGS algorithm. We are among the first to propose and analyze the generalization of the stochastic variance reduced gradient (SVRG) to Riemannian manifolds. As an extension of the linearly convergent VR-PCA, it is significant and nontrivial for the proposed algorithm to theoretically achieve a further speedup and empirically make a difference, due to our respect to the inherent geometry of the problem.

1 INTRODUCTION

The problem of finding a group of top eigenvectors of a symmetric matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ is among the core and long-standing topics in numerical computing (Wilkinson, 1988), and plays fundamental roles in various scientific and engineering computing problems, such as numerical computation (Golub and Van Loan, 1996; Press et al., 2007), structural analysis (Torbjorn Ringertz, 1997), kernel approximation (Drineas and Mahoney, 2005) and spectral clustering (Ng et al., 2002; Xu and Ke, 2016b) in machine learning. Varieties of solvers for this problem have been proposed such as the deterministic power iteration (Golub and Van Loan, 1996) and (block) Lanczos algorithm (Parlett, 1998), randomized SVD (Halko et al., 2011), Oja's stochastic PCA (Oja and Karhunen, 1985), online learning of eigenvectors (Garber et al., 2015), stochastic PCA with variance reduction (Shamir, 2015, 2016), and so on. Studies from the optimization perspective are popular as well, including the trace penalty minimization (Wen et al., 2013) and notably the unconstrained maximization in the Riemannian setting (Edelman et al., 1999; Absil et al., 2008; Wen and Yin, 2013). The Riemannian formulation yields a geodesically non-convex problem:

$$\max_{\mathbf{X}\in \mathrm{St}(n,k)} f(\mathbf{X}) \triangleq (1/2)\mathrm{tr}(\mathbf{X}^\top \mathbf{A}\mathbf{X}), \tag{1}$$

where $\mathrm{St}(n,k) = \{\mathbf{X} \in \mathbb{R}^{n\times k} : \mathbf{X}^\top\mathbf{X} = \mathbf{I}\}$ constitutes a Riemannian manifold that is called the Stiefel manifold. Inspired by the success of the recently proposed stochastic variance reduced gradient (SVRG) method, in this paper, we generalize this technique from the Euclidean space to Riemannian manifolds. First, the general form of the stochastic variance reduced Riemannian gradient (SVRRG) method is proposed in the stochastic Riemannian gradient optimization framework. It then is specialized to Problem (1) and gives rise to our SVRRG-EIGS algorithm. In addition to the use of Riemannian gradients and retraction operations, we need to introduce one new ingredient, i.e., parallel/vector transport, in order to legitimize SVRG in the Riemannian setting. Built upon the analysis of VR-PCA (Shamir, 2015, 2016), we prove the local, eigengap-dependent and linear convergence of the proposed SVRRG-EIGS. In addition, it is shown that SVRRG-EIGS achieves a further speedup both theoretically and empirically on top of the fast VR-

PCA, due to our respect to the Riemannian geometry of the problem. This speedup is significant and nontrivial, because VR-PCA already converges at a rate of a best possible order with first-order methods, i.e., $O(\log \frac{1}{\epsilon})$. The main contributions of the paper are summarized as follows:

- We generalize SVRG to Riemannian manifolds to obtain the general SVRRG method and the SVRRG-EIGS algorithm special for Problem (1).
- We establish the local, eigengap-dependent and linear convergence of the proposed SVRRG-EIGS algorithm for the underlying geodesically non-convex problem. We are among the first to propose and analyze the generalization of SVRG to Riemannian manifolds.
- We show that SVRRG-EIGS improves over VR-PCA in both theory and practice.

2 RELATED WORK

We briefly review those recently proposed and closely related work on the stochastic (Riemannian) eigensolvers and refer readers to cited papers and references theirin for more relevant work.

There has been a large body of work emerging recently on stochastic eigensolvers. Balsubramani et al. (2013) proved the local, eigengap-dependent and sub-linear convergence of Oja's stochastic PCA (Oja and Karhunen, 1985) for the case $k = 1$. Shamir proposed VR-PCA based on Oja's algorithm, and proved its local, eigengap-dependent and linear convergence for $k = 1$ (i.e., vector version) in (Shamir, 2015) and its nontrivial extension to $k \geq 1$ (i.e., block version) in (Shamir, 2016). The analysis of our proposed algorithm SVRRG-EIGS is built upon and meanwhile significantly extends/improves over that of VR-PCA for the block version, as is shown largely in the supplementary material. Garber et al. (2016) proved the global and eigengap-dependent linear convergence of PCA and the global, eigengap-free and sub-linear convergence of PCA, both for the case $k = 1$, by giving a robust analysis of the shift-and-invert preconditioning method to reduce the target problem to a sequence of linear systems and then leveraging the SVRG for the system solver. Jain et al. (2016) provides an eigengap-dependent, linear-time and single-pass streaming algorithm for the case $k = 1$. Ge et al. (2016) proved the global, eigengap-dependent and linear convergence of generalized eigenvalue problem for the case $k = 1$. Allen-Zhu and Li (2016) proposed a fast SVD decomposition with the global, eigengap-free and sub-linear convergence for the case $k > 1$ via recursive calls to the vector version of SVD, i.e., $k = 1$. As we can see, most of these studies focus on the vector case $k = 1$, while we work on the block case $k \geq 1$ in this paper. Note that some work on PCA rely on the special data representation $\mathbf{A} = \mathbf{B}\mathbf{B}^\top$ where $\mathbf{B} \in \mathbb{R}^{n\times m}$ consists of m n-dimensional vector samples, and manipulate $\mathbf{B}$ instead of $\mathbf{A}$ in analysis. Their results might be inapplicable to the case where we are given only $\mathbf{A}$ without $\mathbf{B}$.

Our work falls into the category of the stochastic Riemannian optimization. From this point of view, Bonnabel (2013) proposed the general form of the Riemannian stochastic gradient descent (SGD) method and proved its global and almost sure convergence to Riemannian stationary points. The results naturally apply to the Riemannian SGD solver for Problem (1). However, as we know, any k eigenvectors of $\mathbf{A}$ constitutes a Riemannian stationary point of the problem. Thus, theoretically, they are not necessarily the top k eigenvectors. Xu et al. (2016) proved the local, eigengap-dependent and sub-linear convergence to a globally optimal solution. Recently the variance reduction technique has been generalized to Riemannian manifolds by three groups of researchers in parallel. Zhang et al. (2016) proposed the general Riemannian SVRG for compact manifolds and proved the global, eigengap-free and sub-linear convergence to Riemannian stationary points for geodesically nonconvex problems including Problem (1). However, their results only apply to the case $k = 1$ for our problem, because the gradient dominance property was only proven to hold for this case. For $k > 1$, it is unclear thus far. Moreover, their method is computationally costly, because it is built upon the geodesic based operations and thus needs the computation of matrix exponential on Stiefel manifolds. And the parallel transport they used has no closed forms for Stiefel manifolds when $k > 1$. In contrast, we use a closed-form vector transport. Kasai et al. (2016) proposed the Riemannian SVRG for Grassmann manifolds and proved its local, eigengap-free and linear convergence. However, they require the Riemannian Hessian at the locally optimal solution to be positive definite, which does not hold for Problem (1). And their method incurs the computational issue similarly. For example, in each inner iteration, their method needs to do SVD twice in order to complete three operations, i.e., logarithmic map, parallel transport and exponential map. In contrast, we use their cheap counterparts (i.e., first-order approximation) without SVD and do not need the logarithmic map operation. A preliminary version of this paper is also among the first to propose the Riemannian SVRG (Xu and Ke, 2016a).

3 PRELIMINARIES

Suppose that λ_i represents the i-th largest eigenvalue of $\mathbf{A}$, $\mathbf{\Sigma} = \mathrm{diag}(\lambda_1, \ldots, \lambda_k)$ and $\mathbf{U}$ is the collection of corresponding top k eigenvectors in columns. Then $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{U}^\top + \mathbf{U}_\perp \mathbf{\Sigma}_\perp \mathbf{U}_\perp^\top$, where $\mathbf{\Sigma}_\perp$ is a diagonal matrix with diagonal entries consisting of the other eigenvalues λ_i, $k < i \leq n$ and $\mathbf{U}_\perp$ is the orthogonal complement of $\mathbf{U}$ corresponding to the other eigenvectors. It is easy to see that an arbitrary globally optimal solution of Problem (1) is given by $\mathbf{X}^\star = \mathbf{U}\mathbf{Q}$ where $\mathbf{Q} \in \mathrm{St}(k, k)$ is an orthogonal matrix, under the assumption that the eigengap $\tau = \lambda_k - \lambda_{k+1} > 0$.

3.1 RIEMANNIAN OPTIMIZATION

Given a Riemmanian manifold $\mathcal{M}$, its tangent space at a point $\mathbf{X} \in \mathcal{M}$, denoted as $T_{\mathbf{X}}\mathcal{M}$, is a Euclidean space that is tangential to and locally linearizes $\mathcal{M}$ around this point (Lee, 2012). One update step of the first-order Riemannian optimization on $\mathcal{M}$ can be written as (Absil et al., 2008):

$$\mathbf{X}_{t+1} = R_{\mathbf{X}_t}(\alpha_{t+1}\xi_{\mathbf{X}_t}), \tag{2}$$

where $\xi_{\mathbf{X}_t} \in T_{\mathbf{X}_t}\mathcal{M}$ is called a tangent vector of $\mathcal{M}$ at $\mathbf{X}_t$ and used as the t-th search direction, $\alpha_{t+1} > 0$ is the learning rate (a.k.a. step size), and $R_{\mathbf{X}_t}(\cdot)$ represents the retraction at $\mathbf{X}_t$ that maps a tangent vector $\xi \in T_{\mathbf{X}_t}\mathcal{M}$ to a point on $\mathcal{M}$. Tangent vectors serving as search directions are generally gradient-related. The gradient of a function $f(\mathbf{X})$ on $\mathcal{M}$, denoted as $\mathrm{Grad} f(\mathbf{X})$, depends on the Riemannian metric, which is a family of smoothly varying inner products on tangent spaces, i.e., $\langle \xi, \eta \rangle_{\mathbf{X}}$, where $\xi, \eta \in T_{\mathbf{X}}\mathcal{M}$ for any $\mathbf{X} \in \mathcal{M}$. The Riemannian gradient $\mathrm{Grad} f(\mathbf{X}) \in T_{\mathbf{X}}\mathcal{M}$ is the unique tangent vector that satisfies

$$\langle \mathrm{Grad} f(\mathbf{X}), \xi \rangle_{\mathbf{X}} = Df(\mathbf{X})[\xi] \tag{3}$$

for any $\xi \in T_{\mathbf{X}}\mathcal{M}$, where $Df(\mathbf{X})[\xi]$ represents the directional derivative of $f(\mathbf{X})$ in the tangent direction ξ. Setting $\xi_{\mathbf{X}_t} = \mathrm{Grad} f(\mathbf{X}_t)$ in (2) leads to the Riemannian gradient (RG) ascent method:

$$\mathbf{X}_{t+1} = R_{\mathbf{X}_t}(\alpha_{t+1}\mathrm{Grad} f(\mathbf{X}_t)), \tag{4}$$

while setting $\xi_{\mathbf{X}_t} = G(y_{t+1}, \mathbf{X}_t)$ in (2) gives us the stochastic Riemannian gradient (Bonnabel, 2013) (SRG) ascent method:

$$\mathbf{X}_{t+1} = R_{\mathbf{X}_t}(\alpha_{t+1}G(y_{t+1}, \mathbf{X}_t)), \tag{5}$$

where y_{t+1} is a random variable such that $\mathbb{E}[f(y_{t+1}, \mathbf{X}_t)|\mathbf{X}_t] = f(\mathbf{X}_t)$, and $G(y_{t+1}, \mathbf{X}_t) \in T_{\mathbf{X}_t}\mathcal{M}$ is the stochastic Riemannian gradient such that $\mathbb{E}[G(y_{t+1}, \mathbf{X}_t)|\mathbf{X}_t] = \mathrm{Grad} f(\mathbf{X}_t)$. According to (Bonnabel, 2013), the SRG method can converge globally and almost surely to a stationary point under mild conditions including $\sum_t \alpha_t = \infty$ and $\sum_t \alpha_t^2 < \infty$ (the latter condition implies that $\alpha_t \to 0$ as $t \to \infty$).

3.2 RIEMANNIAN EIGENSOLVER

For Problem (1), note that $\mathrm{St}(n, k)$ is an embedded Riemannian sub-manifold of the Euclidean space $\mathbb{R}^{n \times k}$. With the metric inherited from the embedding space $\mathbb{R}^{n \times k}$, i.e., $\langle \xi, \eta \rangle_{\mathbf{X}} = \mathrm{tr}(\xi^\top \eta)$, and by (3), we can get the Riemannian gradient[1] of $f(\mathbf{X})$ in Problem (1), i.e.,

$$\mathrm{Grad} f(\mathbf{X}) = (\mathbf{I} - \mathbf{X}\mathbf{X}^\top)\mathbf{A}\mathbf{X} \in T_{\mathbf{X}}\mathrm{St}(n, k).$$

The orthogonal projection onto $T_{\mathbf{X}}\mathrm{St}(n, k)$ under this metric is given by

$$P_{\mathbf{X}}(\zeta) = (\mathbf{I} - \mathbf{X}\mathbf{X}^\top)\zeta + \mathbf{X}\mathrm{skew}(\mathbf{X}^\top \zeta) \tag{6}$$

for any $\zeta \in \mathbb{R}^{n \times k}$, where $\mathrm{skew}(H) = (H - H^\top)/2$. We use the polar decomposition based retraction (Absil et al., 2008)

$$R_{\mathbf{X}}(\xi) = (\mathbf{X} + \xi)(\mathbf{I} + \xi^\top \xi)^{-1/2} \tag{7}$$

for any $\xi \in T_{\mathbf{X}}\mathrm{St}(n, k)$. The deployment of (4) and (5) to Problem (1) generates one Riemannian eigensolver denoted as RG-EIGS and one stochastic Riemannian eigensolver denoted as SRG-EIGS, respectively.

4 SVRRG

Recall that SVRG (Johnson and Zhang, 2013) is built on the vanilla stochastic gradient and achieves the variance reduction through constructing control variates in epoches (Wang et al., 2013). Control variates are of stochastic zero-mean and serve to augment and correct stochastic gradients towards the true gradients. Following (Johnson and Zhang, 2013), SVRG reads

$$g_t(\xi_t, w^{(t-1)}, \tilde{w}) = \nabla\psi_{i_t}(w^{(t-1)}) - (\nabla\psi_{i_t}(\tilde{w}) - \nabla P(\tilde{w})),$$

where $\tilde{w}$ is a version of the variable w estimated at the snapshot point after every m SGD steps, and $\nabla P(\tilde{w}) = \frac{1}{n}\sum_{i=1}^n \nabla\psi_i(\tilde{w})$ is the full gradient at $\tilde{w}$.

Our goal here is to develop the Riemannian counterpart of SVRG, termed as SVRRG and denoted as

[1] Due to the symmetry of $\mathbf{A}$, Riemannian gradients under the Euclidean metric and the canonical metric are the same (Wen and Yin, 2013). However, since the orthogonal projector used in the sequel requires the metrics for the embedded Riemannian sub-manifold and the embedding space to be the same, we choose the Euclidean metric here.

$G(y_{t+1}, \mathbf{X}_t, \tilde{\mathbf{X}})$. One naive extension of SVRG to Riemannian manifolds by substituting Riemannian gradients only can be written as

$$\begin{aligned} & G(y_{t+1}, \mathbf{X}_t, \tilde{\mathbf{X}}) \\ = \; & G(y_{t+1}, \mathbf{X}_t) - (G(y_{t+1}, \tilde{\mathbf{X}}) - \mathrm{Grad} f(\tilde{\mathbf{X}})), \end{aligned}$$

where $G(y_{t+1}, \tilde{\mathbf{X}})$, $\mathrm{Grad} f(\tilde{\mathbf{X}}) \in T_{\tilde{\mathbf{X}}}\mathcal{M}$ and $G(y_{t+1}, \mathbf{X}_t) \in T_{\mathbf{X}_t}\mathcal{M}$. However, this is unsound theoretically, as the stochastic Riemannian gradient $G(y_{t+1}, \mathbf{X}_t)$ and the control variate $G(y_{t+1}, \tilde{\mathbf{X}}) - \mathrm{Grad} f(\tilde{\mathbf{X}})$ reside in two different tangent spaces and it hence renders their difference $G(y_{t+1}, \mathbf{X}_t, \tilde{\mathbf{X}})$ not well-defined in a Riemannian space. To rectify this issue, we need the operation that can move tangent vectors from one point to another along the geodesics in parallel, namely parallel transport. Our control variate thus need be parallel transported from $\tilde{\mathbf{X}}$ to $\mathbf{X}_t$. For computational efficiency, vector transport as its first-order approximation is often used in practice (Absil et al., 2008).

Vector transport of a tangent vector from point $\tilde{\mathbf{X}}$ to point $\mathbf{X}_t$, denoted as $\mathcal{T}_{\tilde{\mathbf{X}} \to \mathbf{X}_t}$, is a mapping from the tangent space $T_{\tilde{\mathbf{X}}}\mathcal{M}$ to the tangent space $T_{\mathbf{X}_t}\mathcal{M}$. When $\mathcal{M}$ is an embedded Riemannian sub-manifold of a Euclidean space, it can be simply defined as $\mathcal{T}_{\tilde{\mathbf{X}} \to \mathbf{X}_t}(\xi_{\tilde{\mathbf{X}}}) = P_{\mathbf{X}_t}(\xi_{\tilde{\mathbf{X}}})$. With vector transport, we now have a well-defined SVRRG in $T_{\mathbf{X}_t}\mathcal{M}$, i.e.,

$$\begin{aligned} & G(y_{t+1}, \mathbf{X}_t, \tilde{\mathbf{X}}) \\ = & G(y_{t+1}, \mathbf{X}_t) - \mathcal{T}_{\tilde{\mathbf{X}} \to \mathbf{X}_t}(G(y_{t+1}, \tilde{\mathbf{X}}) - \mathrm{Grad} f(\tilde{\mathbf{X}})). \end{aligned}$$

The use of Riemannian gradients and vector transport makes SVRRG significantly different from SVRG. We then arrive at the SVRRG method:

$$\mathbf{X}_{t+1} = R_{\mathbf{X}_t}(\alpha_{t+1} G(y_{t+1}, \mathbf{X}_t, \tilde{\mathbf{X}})). \tag{8}$$

Note that the SVRRG method is naturally subsumed into the SRG method (5), and thus enjoys all its properties including the almost sure convergence.

Algorithm 1 SVRRG

Require: $\mathbf{A}$, $\tilde{\mathbf{X}}_0$, α, epoch length m

```
for s = 1, 2, ··· do
    Compute Grad f(X̃_{s-1})
    X_0 = X̃_{s-1}
    for t = 1, 2, ··· , m do
        Pick y_t ∈ 𝒴 uniformly at random
        Compute X_t = R_{X_{t-1}}(α G(y_t, X_{t-1}, X̃_{s-1}))
    end for
    X̃_s = X_m
end for
```

5 SVRRG-EIGS

In this section, we deploy the SVRRG method to Problem (1), then giving rise to a new eigensolver termed as SVRRG-EIGS. Assume that $\mathbf{A} = \frac{1}{L}\sum_{i=1}^{L} \tilde{\mathbf{A}}_i$, y is a random variable taking values in $\mathcal{Y} = \{1, 2, \cdots, L\}$ and $\mathbf{A}_{t+1} = \tilde{\mathbf{A}}_{y_{t+1}}$. In practice, $\tilde{\mathbf{A}}_i$ can be obtained from a partitioning of $\mathbf{A}$. We then can write the stochastic gradient and control variate as

$$\begin{aligned} G(y_{t+1}, \mathbf{X}_t) &= (\mathbf{I} - \mathbf{X}_t\mathbf{X}_t^\top)\mathbf{A}_{t+1}\mathbf{X}_t, \\ G(y_{t+1}, \tilde{\mathbf{X}}) - \mathrm{Grad} f(\tilde{\mathbf{X}}) &= (\mathbf{I} - \tilde{\mathbf{X}}\tilde{\mathbf{X}}^\top)(\mathbf{A}_{t+1} - \mathbf{A})\tilde{\mathbf{X}}, \end{aligned}$$

respectively. Using the orthogonal projector (6), the transported control variate can be written as

$$\begin{aligned} & \mathcal{T}_{\tilde{\mathbf{X}} \to \mathbf{X}_t}(G(y_{t+1}, \tilde{\mathbf{X}}) - \mathrm{Grad} f(\tilde{\mathbf{X}})) \\ = \; & (\mathbf{I} - \mathbf{X}_t\mathbf{X}_t^\top)(\mathbf{I} - \tilde{\mathbf{X}}\tilde{\mathbf{X}}^\top)(\mathbf{A}_{t+1} - \mathbf{A})\tilde{\mathbf{X}} \\ & + \mathbf{X}_t \mathrm{skew}(\mathbf{X}_t^\top(\mathbf{I} - \tilde{\mathbf{X}}\tilde{\mathbf{X}}^\top)(\mathbf{A}_{t+1} - \mathbf{A})\tilde{\mathbf{X}}) \\ = \; & (\mathbf{I} - \mathbf{X}_t\mathbf{X}_t^\top)(\mathbf{A}_{t+1} - \mathbf{A})\tilde{\mathbf{X}} \\ & - (\mathbf{I} - \mathbf{X}_t\mathbf{X}_t^\top)\tilde{\mathbf{X}}\tilde{\mathbf{X}}^\top(\mathbf{A}_{t+1} - \mathbf{A})\tilde{\mathbf{X}} \\ & + \mathbf{X}_t \mathrm{skew}(\mathbf{X}_t^\top(\mathbf{I} - \tilde{\mathbf{X}}\tilde{\mathbf{X}}^\top)(\mathbf{A}_{t+1} - \mathbf{A})\tilde{\mathbf{X}}). \end{aligned}$$

SVRRG then reads

$$\begin{aligned} & G(y_{t+1}, \mathbf{X}_t, \tilde{\mathbf{X}}) \\ = \; & (\mathbf{I} - \mathbf{X}_t\mathbf{X}_t^\top)\mathbf{A}_{t+1}\mathbf{X}_t \\ & - \mathcal{T}_{\tilde{\mathbf{X}} \to \mathbf{X}_t}(G(y_{t+1}, \tilde{\mathbf{X}}) - \mathrm{Grad} f(\tilde{\mathbf{X}})) \\ = \; & (\mathbf{I} - \mathbf{X}_t\mathbf{X}_t^\top)\mathbf{A}\mathbf{X}_t \\ & + (\mathbf{I} - \mathbf{X}_t\mathbf{X}_t^\top)(\mathbf{A}_{t+1} - \mathbf{A})(\mathbf{X}_t - \tilde{\mathbf{X}}) \\ & + (\mathbf{I} - \mathbf{X}_t\mathbf{X}_t^\top)\tilde{\mathbf{X}}\tilde{\mathbf{X}}^\top(\mathbf{A}_{t+1} - \mathbf{A})\tilde{\mathbf{X}} \\ & - \mathbf{X}_t \mathrm{skew}(\mathbf{X}_t^\top(\mathbf{I} - \tilde{\mathbf{X}}\tilde{\mathbf{X}}^\top)(\mathbf{A}_{t+1} - \mathbf{A})\tilde{\mathbf{X}}) \\ \triangleq \; & \mathrm{Grad} f(\mathbf{X}_t) + \mathbf{W}_t, \end{aligned}$$

where $\mathbf{W}_t \in T_{\mathbf{X}_t}\mathrm{St}(n, k)$ is a stochastic zero-mean term conditioned on $\mathbf{X}_t$. As we see above, SVRRG has significantly extended SVRG used in VR-PCA. Although the factor $(\mathbf{X}_t - \tilde{\mathbf{X}})$ present in $\mathbf{W}_t$ works well empirically, in order for ease of theoretical analysis we follow (Shamir, 2016) to replace $\tilde{\mathbf{X}}$ with $\tilde{\mathbf{X}}\mathbf{Q}_t$, where $\mathbf{Q}_t = \mathbf{P}_2\mathbf{P}_1^\top$ and $\mathbf{X}_t^\top\tilde{\mathbf{X}} = \mathbf{P}_1\mathbf{\Lambda}\mathbf{P}_2^\top$ is the SVD of $\mathbf{X}_t^\top\tilde{\mathbf{X}}$. With a bit abuse of the notation for $\mathbf{W}$, we have the final SVRRG for Problem (1) written as,

$$\begin{aligned} & G(y_{t+1}, \mathbf{X}_t, \tilde{\mathbf{X}}, \mathbf{Q}_t) \\ = \; & (\mathbf{I} - \mathbf{X}_t\mathbf{X}_t^\top)\mathbf{A}\mathbf{X}_t \\ & + (\mathbf{I} - \mathbf{X}_t\mathbf{X}_t^\top)(\mathbf{A}_{t+1} - \mathbf{A})(\mathbf{X}_t - \tilde{\mathbf{X}}\mathbf{Q}_t) \\ & + (\mathbf{I} - \mathbf{X}_t\mathbf{X}_t^\top)\tilde{\mathbf{X}}\tilde{\mathbf{X}}^\top(\mathbf{A}_{t+1} - \mathbf{A})\tilde{\mathbf{X}}\mathbf{Q}_t \\ & - \mathbf{X}_t \mathrm{skew}(\mathbf{X}_t^\top(\mathbf{I} - \tilde{\mathbf{X}}\tilde{\mathbf{X}}^\top)(\mathbf{A}_{t+1} - \mathbf{A})\tilde{\mathbf{X}}\mathbf{Q}_t) \\ \triangleq \; & \mathrm{Grad} f(\mathbf{X}_t) + \mathbf{W}_t. \end{aligned}$$

Thus the intermediate update in the tangent space is

$$\text{SVRRG-EIGS}: \mathbf{X}_t + \alpha_{t+1}\text{Grad}f(\mathbf{X}_t) + \alpha_{t+1}\mathbf{W}_t.$$

For comparison, the counterparts for RG-EIGS and SRG-EIGS are given similarly as follows

$$\begin{aligned}\text{RG-EIGS}: &\quad \mathbf{X}_t + \alpha_{t+1}\text{Grad}f(\mathbf{X}_t),\\ \text{SRG-EIGS}: &\quad \mathbf{X}_t + \alpha_{t+1}\text{Grad}f(\mathbf{X}_t)\\ &\quad +\alpha_{t+1}(\mathbf{I}-\mathbf{X}_t\mathbf{X}_t^\top)(\mathbf{A}_{t+1}-\mathbf{A})\mathbf{X}_t.\end{aligned}$$

We can see that intermediate steps in both SRG-EIGS and SVRRG-EIGS amount to moving forward along the Riemannian gradient direction that is perturbed by a stochastic zero-mean noise term in the tangent space. However, the stochastic zero-mean term with SRG-EIGS, i.e., $(\mathbf{I}-\mathbf{X}_t\mathbf{X}_t^\top)(\mathbf{A}_{t+1}-\mathbf{A})\mathbf{X}_t$, always carries a constant variance. It thus needs a diminishing learning rate α_t in order to reduce the variance and to ensure the convergence. As a result, the convergence rate is compromised. On the contrary, SVRRG-EIGS keeps boosting the variance reduction of the stochastic zero-mean term $\mathbf{W}_t$ during iterations. The variance of $\mathbf{W}_t$ is dominated by quantities $\|\mathbf{X}_t-\tilde{\mathbf{X}}\mathbf{Q}_t\|$, $\|(\mathbf{I}-\mathbf{X}_t\mathbf{X}_t^\top)\tilde{\mathbf{X}}\mathbf{Q}_t\|$ and $\|\mathbf{X}_t^\top(\mathbf{I}-\tilde{\mathbf{X}}\tilde{\mathbf{X}}^\top)\|$. Instead of learning rate, these quantities repeatedly decay till vanishing, as $\mathbf{X}_t$ and $\tilde{\mathbf{X}}\mathbf{Q}_t$ become increasingly close to each other. Since it ensures a decaying variance without learning rate involved, SVRRG-EIGS is able to use a fixed learning rate $\alpha_t=\alpha$ and thus to achieve a much faster convergence rate.

The algorithmic steps of SVRRG-EIGS are given in Algorithm 2.

Algorithm 2 SVRRG-EIGS

Require: $\mathbf{A}$, $\tilde{\mathbf{X}}_0$, α, m

for $s=1,2,\cdots$ **do**
  $\tilde{\mathbf{G}}^{s-1}=(\mathbf{I}-\tilde{\mathbf{X}}_{s-1}\tilde{\mathbf{X}}_{s-1}^\top)\mathbf{A}\tilde{\mathbf{X}}_{s-1}$
  $\mathbf{X}_0=\tilde{\mathbf{X}}_{s-1}$
  for $t=1,2,\cdots,m$ **do**
    Pick $y_t\in\{1,2,\cdots,L\}$ uniformly at random
    $\tilde{\mathbf{G}}_{t-1}^{s-1}=(\mathbf{I}-\tilde{\mathbf{X}}_{s-1}\tilde{\mathbf{X}}_{s-1}^\top)\mathbf{A}_t\tilde{\mathbf{X}}_{s-1}$
    $\mathbf{X}_{t-1}^\top\tilde{\mathbf{X}}_{s-1}\overset{\text{svd}}{=}\mathbf{P}_1\mathbf{\Lambda}\mathbf{P}_2^\top$
    $\Delta_{t-1}^{s-1}=(\tilde{\mathbf{G}}_{t-1}^{s-1}-\tilde{\mathbf{G}}^{s-1})\mathbf{P}_2\mathbf{P}_1^\top$
    $\mathcal{T}(\Delta_{t-1}^{s-1})=\Delta_{t-1}^{s-1}-\mathbf{X}_{t-1}\text{sym}(\Delta_{t-1}^{s-1})$
    $\mathbf{G}_{t-1}=(\mathbf{I}-\mathbf{X}_{t-1}\mathbf{X}_{t-1}^\top)\mathbf{A}_t\mathbf{X}_{t-1}$
    $\hat{\mathbf{G}}_{t-1}^{s-1}=\mathbf{G}_{t-1}-\mathcal{T}(\Delta_{t-1}^{s-1})$
    $\mathbf{Y}_t=\mathbf{X}_{t-1}+\alpha\hat{\mathbf{G}}_{t-1}^{s-1}$
    $\mathbf{X}_t=\mathbf{Y}_t(\mathbf{Y}_t^\top\mathbf{Y}_t)^{-1/2}$
  end for
  $\tilde{\mathbf{X}}_s=\mathbf{X}_m$
end for

5.1 THEORETICAL ANALYSIS

We now study theoretical properties of SVRRG-EIGS, which is built upon the analysis of VR-PCA in (Shamir, 2016). The potential function is defined as $\Theta_t\triangleq\Theta(\mathbf{X}_t,\mathbf{U})=k-\|\mathbf{U}^\top\mathbf{X}_t\|_F^2$. It is easy to see that $\Theta_t\in[0,k]$ and $\Theta_t=0$ if and only if $\mathbf{X}_t=\mathbf{U}\mathbf{Q}$, where $\mathbf{Q}\in\text{St}(k,k)$. Note that the only assumptions we make in the analysis are that the eigengap $\tau>0$ and $\Theta_0<\delta$, where $\delta\in(0,1)$. In what follows, we use a fixed learning rate α. In addition, let $\beta=\|\mathbf{A}\|_2$, $\tilde{\Theta}_s=\Theta(\tilde{\mathbf{X}}_s,\mathbf{U})$ and $\mathbf{Y}_t=\mathbf{X}_t+\alpha\text{Grad}f(\mathbf{X}_t)$. All the proofs of lemmas herein for the analysis are provided in the supplementary material.

First, we can write

$$\begin{aligned}&\|\mathbf{X}_{t+1}^\top\mathbf{U}\|_F^2\\ =\;&\|R_{\mathbf{X}_t}(\alpha G(y_{t+1},\mathbf{X}_t,\tilde{\mathbf{X}},\mathbf{Q}_t))^\top\mathbf{U}\|_F^2\\ =\;&\text{tr}((a_1(\mathbf{X}_t)+b_1(\mathbf{W}_t))(a_2(\mathbf{X}_t)+b_2(\mathbf{W}_t))^{-1}),\end{aligned}$$

where[2]

$$\begin{aligned}a_1(\mathbf{X}_t)&=\mathbf{Y}_t^\top\mathbf{U}\mathbf{U}^\top\mathbf{Y}_t+\alpha^2\mathbf{W}_t^\top\mathbf{U}\mathbf{U}^\top\mathbf{W}_t,\\ b_1(\mathbf{W}_t)&=2\alpha\,\text{sym}(\mathbf{Y}_t^\top\mathbf{U}\mathbf{U}^\top\mathbf{W}_t),\\ a_2(\mathbf{X}_t)&=\mathbf{Y}_t^\top\mathbf{Y}_t+\alpha^2\mathbf{W}_t^\top\mathbf{W}_t,\\ b_2(\mathbf{W}_t)&=2\alpha\,\text{sym}(\mathbf{Y}_t^\top\mathbf{W}_t).\end{aligned}$$

Note that $\mathbb{E}[b_1(\mathbf{W}_t)|\mathbf{X}_t]=\mathbb{E}[b_2(\mathbf{W}_t)|\mathbf{X}_t]=\mathbf{0}$, $a_1(\mathbf{X}_t)\succcurlyeq\mathbf{Y}_t^\top\mathbf{U}\mathbf{U}^\top\mathbf{Y}_t$ and $a_2(\mathbf{X}_t)\preccurlyeq\mathbf{Y}_t^\top\mathbf{Y}_t+\alpha^2\nu_t^2\mathbf{I}$ due to the following lemma.

Lemma 5.1. *Let* $\tilde{\beta}=\max_i\|\tilde{\mathbf{A}}_i-\mathbf{A}\|_2$. *Then*

$$\begin{aligned}\|\mathbf{W}_t\|_2&\le\mu\triangleq 4\tilde{\beta},\\ \|\mathbf{W}_t\|_F^2&\le\nu_t^2\triangleq 24\tilde{\beta}^2(\Theta_t+\tilde{\Theta}_{s-1}).\end{aligned}$$

Note that ν_t is a constant conditioned on $\mathbf{X}_t$. Despite a more complicated stochastic zero-mean term $\mathbf{W}_t$, we manage to bound it by a quantity about $\Theta_t+\tilde{\Theta}_{s-1}$ which will help the subsequent analysis.

Lemma 5.2. *If* $\mathbf{C}_1\succcurlyeq\mathbf{0}$, $\mathbf{C}_2\succcurlyeq\mathbf{0}$ *and* $\mathbf{D}_2\succcurlyeq\mathbf{D}_1\succ\mathbf{0}$, *then* $\text{tr}(\mathbf{C}_1\mathbf{D}_1^{-1})\ge\text{tr}((\mathbf{C}_1-\mathbf{C}_2)\mathbf{D}_2^{-1})$.

By Lemma 5.2 and abusing notations for $a_i(\mathbf{X}_t)$, $i=1,2$, we get

$$\begin{aligned}&\|\mathbf{X}_{t+1}^\top\mathbf{U}\|_F^2\\ \ge&\text{tr}((a_1(\mathbf{X}_t)+b_1(\mathbf{W}_t))(a_2(\mathbf{X}_t)+b_2(\mathbf{W}_t))^{-1}),\end{aligned}$$

where $a_1(\mathbf{X}_t)=\mathbf{Y}_t^\top\mathbf{U}\mathbf{U}^\top\mathbf{Y}_t$ and $a_2(\mathbf{X}_t)=\mathbf{Y}_t^\top\mathbf{Y}_t+\alpha^2\nu_t^2\mathbf{I}$ now. And note that $a_i(\mathbf{X}_t)$, $i=1,2$, become deterministic conditioned on $\mathbf{X}_t$.

[2] $\text{sym}(H)=\frac{1}{2}(H+H^\top)$.

Lemma 5.3. *For $i = 1, 2$ and any $\varsigma \in [0, 1]$,*

$$\|b_i(\mathbf{W}_t)\|_F \leq 2(1+\alpha\beta)\alpha\nu,$$
$$\|a_1(\mathbf{X}_t)\|_2 + \|b_1(\mathbf{W}_t)\|_2 \leq (1+\alpha\beta)\left(1+\alpha\left(\beta+2\mu\right)\right),$$
$$a_2(\mathbf{X}_t) + \varsigma b_2(\mathbf{W}_t) \succcurlyeq \left(1-2(1+\alpha\beta)\alpha\mu\right)\mathbf{I}.$$

By Lemma 5 in (Shamir, 2016) and Lemma 5.3 above, we can get

$$\mathbb{E}[\|\mathbf{X}_{t+1}^\top \mathbf{U}\|_F^2 | \mathbf{X}_t] \geq \mathrm{tr}(a_1(\mathbf{X}_t)a_2^{-1}(\mathbf{X}_t)) - \eta\alpha^2\nu_t^2,$$

where $(1+\alpha\beta)\alpha\mu < \frac{1}{2}$ and $\eta = \frac{4(1+\alpha\beta)^2(2+(1+\alpha\beta)^2)}{(1-2(1+\alpha\beta)\alpha\mu)^3}$.

Lemma 5.4. $\mathrm{tr}(\mathbf{X}_t^\top\mathbf{U}\mathbf{U}^\top(\mathbf{I} - \mathbf{X}_t\mathbf{X}_t^\top)\mathbf{A}\mathbf{X}) \geq \tau(\|\mathbf{X}_t^\top\mathbf{U}\|_F^2 - \|\mathbf{X}_t^\top\mathbf{U}\mathbf{U}^\top\mathbf{X}_t\|_F^2)$.

This lemma is the key to the improvement brought by our algorithm over the VR-PCA as we will see shortly, and is used in proving the following lemma.

Lemma 5.5. *If $\alpha^2(\beta^2 + 48k\tilde{\beta}^2) < 1$, then*

$$\begin{aligned}&\mathrm{tr}(a_1(\mathbf{X}_t)a_2^{-1}(\mathbf{X}_t))\\ \geq\ & 2\alpha\tau(\|\mathbf{X}_t^\top\mathbf{U}\|_F^2 - \|\mathbf{U}^\top\mathbf{X}_t\mathbf{X}_t^\top\mathbf{U}\|_F^2)\\ &+ \|\mathbf{X}_t^\top\mathbf{U}\|_F^2 - \alpha^2(1+2\alpha\beta)(2\beta^2\Theta_t + k\nu_t^2).\end{aligned}$$

Note that the proof of Lemma 5.5 is different from (Shamir, 2016). See the supplementary material for details.

If $\min\{2(1+\alpha\beta)\alpha\mu, \alpha^2(\beta^2+48k\tilde{\beta}^2)\} < 1$, then by Lemma 5.5 above and Lemma 7 in (Shamir, 2016) we can arrive at

$$\begin{aligned}\mathbb{E}[\Theta_{t+1}|\mathbf{X}_t] \leq&(1-\alpha(\frac{2}{k}\tau\sigma_t^2\|\mathbf{X}_t^\top\mathbf{U}\|_F^2 - \gamma\alpha))\Theta_t\\ &+ \rho\tilde{\Theta}_{s-1},\end{aligned}$$

where $\sigma_t \triangleq \sigma_{\min}(\mathbf{X}_t^\top\mathbf{U})$ represents the smallest singular value of $\mathbf{X}_t^\top\mathbf{U}$, $\gamma = (1+2\alpha\beta)(2\beta^2+24k\tilde{\beta}^2) + 24\eta\tilde{\beta}^2$ and $\rho = 24(k(1+2\alpha\beta)+\eta)\alpha^2\tilde{\beta}^2$. It is worth noting that the coefficient for the dominating first-order term about α above is $\frac{2}{k}\tau\sigma_t^2\|\mathbf{X}_t^\top\mathbf{U}\|_F^2$, while the counterpart[3] in (Shamir, 2016) is $\frac{4/5}{k}\tau\sigma_t^2\|\mathbf{X}_t^\top\mathbf{U}\|_F^2$. We thus achieve an improvement by about $\frac{6/5}{k}\tau\sigma_t^2\|\mathbf{X}_t^\top\mathbf{U}\|_F^2$ for each single iteration, and it increases with iterations because both σ_t and $\|\mathbf{X}_t^\top\mathbf{U}\|_F$ keep increasing. For a small α that is often the case in our context, this improvement is indeed negligible. However, that is only for one single iteration. With a large number of iterations which is also often the case in our context, the accumulation of such small improvements can no longer be negligible and will make a visible difference as observed in our experiments.

[3]This is the most appropriate time for the comparison because Shamir (2016) uses implicit constants afterwards.

If $\|\mathbf{X}_t^\top\mathbf{U}\|_F^2 > k-\delta$, i.e., $\Theta_t < \delta$, then $\sigma_t^2 > 1-\delta$ otherwise we get the contradiction that $\|\mathbf{X}_t^\top\mathbf{U}\|_F^2 \leq k - 1 + \sigma_t^2 \leq k-1+1-\delta = k-\delta$. In this case, we get

$$\begin{aligned}\mathbb{E}[\Theta_{t+1}] =&\mathbb{E}[\mathbb{E}[\Theta_{t+1}|\mathbf{X}_t]]\\ \leq&(1-\alpha(2\xi\tau-\gamma\alpha))\mathbb{E}[\Theta_t] + \rho\mathbb{E}[\tilde{\Theta}_{s-1}], \quad (9)\end{aligned}$$

where $\xi = 1 - \frac{(k+1-\delta)\delta}{k}$.

In order for the recurrence of the above inequality, we need that $\Theta_t < \delta$ holds for a number of consecutive iterations with high probability, given that $\tilde{\Theta}_{s-1} < \delta$. We can make it by following the concentration of martingale argument. In fact, based on the above inequality, we can construct a super-martingale with bounded differences. By the Azuma-Hoeffding inequality we then can get the following lemma,

Lemma 5.6. *For any $\tilde{\iota} \in (0,1)$, if α and m satisfy that $\alpha < 2\xi\tau/\gamma$, $\min\{2(1+\alpha\beta)\alpha\mu, \alpha^2(\beta^2+48k\tilde{\beta}^2)\} < 1$ and $\tilde{\Theta}_{s-1} + km\rho + \theta\sqrt{2m\log(1/\iota)} < \delta$, then $\Theta_t < \delta$ holds for all $t = 1, 2, \cdots, m$ with probability at least $1-\tilde{\iota}$, where $\theta = \frac{4k(\beta+4\tilde{\beta})\alpha}{1-(\beta+4\tilde{\beta})\alpha} + k\rho$.*

Note that given a fixed m, α that satisfies $\tilde{\Theta}_{s-1} + km\rho + \theta\sqrt{2m\log(1/\iota)} < \delta$ depends on the epoch. Thus it is denoted as $\alpha_s(m)$. Let $\kappa = 1-\alpha(2\xi\tau-\gamma\alpha)$. Under the conditions of Lemma 5.6, we now can call (9) recursively and get

$$\begin{aligned}&\mathbb{E}[\tilde{\Theta}_{s+1}]\\ =&\mathbb{E}[\Theta_m] \leq \kappa\mathbb{E}[\Theta_{m-1}] + \rho\mathbb{E}[\tilde{\Theta}_s]\\ \leq&\kappa^m\mathbb{E}[\Theta_0] + \rho\sum_{i=0}^{m-1}\kappa^i\mathbb{E}[\tilde{\Theta}_s]\\ =&(\kappa^m + \rho\sum_{i=0}^{m-1}\kappa^i)\mathbb{E}[\tilde{\Theta}_s] \leq (\kappa^m + \frac{\rho}{1-\kappa})\mathbb{E}[\tilde{\Theta}_s]\\ \leq&\left(\exp\{-m\alpha(2\xi\tau-\gamma\alpha)\} + \frac{\rho}{1-\kappa}\right)\mathbb{E}[\tilde{\Theta}_s].\end{aligned}$$

For any $\omega \in (0, \frac{1}{2})$, when[4]

$$\alpha < \min\{(\beta^2+48k\tilde{\beta}^2)^{-1/2}, \frac{1}{4\mu}, \frac{2\xi\tau}{\gamma}, \alpha_s(m)\}$$

and additionally

$$\exp\{-m\alpha(2\xi\tau-\gamma\alpha)\} < \omega, \quad \frac{\rho}{1-\kappa} < 1-2\omega,$$

namely $\alpha < \frac{2(1-2\omega)\xi\tau}{24(3k+\eta)\tilde{\beta}+(1-2\omega)\gamma} < \frac{2\xi\tau}{\gamma}$ and $m > \frac{-\log\omega}{\alpha(2\xi\tau-\gamma\alpha)}$, we have $\mathbb{E}[\tilde{\Theta}_{s+1}] \leq (1-\omega)\mathbb{E}[\tilde{\Theta}_s]$. Then,

[4]If $\alpha^2(\beta^2+48k\tilde{\beta}^2) < 1$, then $\alpha\beta < 1$ and thus $2(1+\alpha\beta)\alpha\mu < 4\alpha\mu$. If $4\alpha\mu < 1$ then $2(1+\alpha\beta)\alpha\mu < 1$.

$\mathbb{E}[\tilde{\Theta}_S] \leq (1-\omega)^S \tilde{\Theta}_0$ with probability at least $1 - S\tilde{\iota}$ if we further have $\alpha < \min_{1\leq s\leq S} \alpha_s(m)$, where the number of epoches S for an ϵ-accurate solution is set such that $(1-\omega)^S \leq \epsilon$, i.e., $S = O(\frac{1}{1-\omega}\log\frac{1}{\epsilon})$, corresponding to the number of iterations $T = mS = O(\frac{1}{\xi^2\tau^2}\log\frac{1}{\epsilon})$. If we set $\tilde{\iota} \in (0, \frac{\iota}{S})$ where $\iota \in (0,1)$, then we get an ϵ-accurate solution with probability at least $1-\iota$.

Therefore, we arrive at the following theorem.

Theorem 5.7. *Given $\mathbf{A} \in \mathbb{R}^{n\times n}$ with $\mathbf{A}^\top = \mathbf{A}$, for any $\iota \in (0,1)$, $\delta \in (0,1)$ and $\omega \in (0, \frac{1}{2})$, if $\tau > 0$ and $\tilde{\Theta}_0 < \delta$, then with probability at least $1-\iota$ the SVRRG-EIGS algorithm is able to reach a globally ϵ-optimal solution by running $T = O(\frac{1}{\xi^2\tau^2}\log\frac{1}{\epsilon})$ iterations in expectation, provided that $0 < \alpha < \min\{(\beta^2 + 48k\tilde{\beta}^2)^{-\frac{1}{2}}, \frac{1}{4\mu}, \frac{2(1-2\omega)\xi\tau}{24(3k+\eta)\tilde{\beta}+(1-2\omega)\gamma}, \min_{1\leq s\leq S}\alpha_s(m)\}$ and $m > \frac{-\log\omega}{\alpha(2\xi\tau-\gamma\alpha)}$, where $S = O(\frac{1}{1-\omega}\log\frac{1}{\epsilon})$.*

Note that if the constant factor with ξ is considered, SVRRG-EIGS and VR-PCA actually have the iteration complexity $O(\frac{1}{4\xi^2\tau^2}\log\frac{1}{\epsilon})$ and $O(\frac{1}{\frac{9}{25}\xi^2\tau^2}\log\frac{1}{\epsilon})$, respectively. Thus, SVRRG-EIGS achieves the linear convergence[5] at an improved rate over VR-PCA.

Last, suppose that the single column sampling of $\mathbf{A}$ is used. Then each epoch will have the complexity equal to $O(k\mathrm{nnz}(\mathbf{A}) + nk^2)$ from computing a full gradient plus $O(m(k\frac{\mathrm{nnz}(\mathbf{A})}{n} + nk^2))$ for the inner loop in the amortised sense. Thus, the total complexity of Algorithm 2 is $O(k(\mathrm{nnz}(\mathbf{A}) + nk + (\frac{\mathrm{nnz}(\mathbf{A})}{n} + nk)\frac{1}{\xi^2\tau^2})\log\frac{1}{\epsilon})$ for an ϵ-optimal solution.

6 EXPERIMENTS

In this section, we empirically verify the linear convergence rate of the SVRRG-EIGS algorithm and compare its performance with that of RG-EIGS, SRG-EIGS and VR-PCA. Among various implementations of the RG-EIGS with different choices of a Riemannian metric and a retraction in (2), we choose the one with the canonical metric and the Cayley transformation based retraction (Wen and Yin, 2013), because it is frequently cited and its code is publicly available[6]. This version of the RG-EIGS uses a non-monotone line search with the well-known Barzilai-Borwein step size, which significantly reduces the iteration number and performs well in practice. We adapt the code of VR-PCA provided by the author to our case, because it was originally designed to handle vectorial data[7]. All the three solvers, RG-EIGS, SRG-EIGS and VR-PCA, are fed with the same random initial value of $\mathbf{X}$, where each entry is sampled from the standard normal distribution $\mathcal{N}(0,1)$ and then all entries as a whole are orthogonalized. SRG-EIGS uses the decaying learning rate $\alpha_t = \frac{c}{t}$, where c will be tuned. In order to generate an initial point $\tilde{\mathbf{X}}_0$ with $\tilde{\Theta}_0 < \delta$ for both SVRRG-EIGS and VR-PCA, we use the SRG-EIGS to produce a low-precision solution as $\tilde{\Theta}_0$. But note that other types of solvers are also applicable such as deterministic solvers (Wen and Yin, 2013) and doubly stochastic solvers (Xu et al., 2016).

Different solvers are tested on a real and symmetric matrix, **Schenk**[8], of size $10,728 \times 10,728$ having $85,000$ nonzero entries. It is partitioned into column blocks of block size $10,724 \times 100$ such that $\mathbf{A} = \frac{1}{L}\sum_{i=1}^{L}\tilde{\mathbf{A}}_i$ where $L = \lceil\frac{10728}{100}\rceil$ and the i-th column block in each $\tilde{\mathbf{A}}_i$ is equal to that of $L\mathbf{A}$ and all others are zero blocks. We set $k = 3$. Both VR-PCA and SVRRG-EIGS are able to use a fixed learning rate $\alpha = \frac{d}{\|\mathbf{A}\|_1\sqrt{n}}$, where d will be tuned and $\|\mathbf{A}\|_1$ represents the matrix 1-norm. The best-tuned values of d for VR-PCA and SVRRG-EIGS are $d = 21.50$ and $d = 4.06$ in the following experiments, respectively. The epoch length is set to $m = \frac{1}{2}L$, i.e., each epoch takes 1.5 passes over $\mathbf{A}$ (including one pass for computing the full gradient). Accordingly, the epoch length of SRG-EIGS is set to $m = \frac{3}{2}L$. In addition, we set $\mathbf{Q}_t = \mathbf{I}$.

The performance of algorithms is evaluated using four quality measures: feasibility $\|\mathbf{X}^\top\mathbf{X} - \mathbf{I}\|_F$, relative error function $E(\mathbf{X}) \triangleq 1 - \frac{\mathrm{tr}(\mathbf{X}^\top\mathbf{A}\mathbf{X})}{\max_{\mathbf{X}\in\mathrm{St}(n,k)}\mathrm{tr}(\mathbf{X}^\top\mathbf{A}\mathbf{X})}$, normalized potential function $\mathbf{\Theta}(\mathbf{X})/k = 1 - \frac{\|\mathbf{X}^\top\mathbf{U}\|_F^2}{k}$ and the potential function used in (Xu et al., 2016) that is defined as $1 - \cos^2\angle_{\max}(\mathbf{U},\mathbf{X})$, where $\angle_{\max}(\mathbf{U},\mathbf{X})$ represents the maximal principal angle between $\mathbf{U}$ and $\mathbf{X}$. The ground truths in these measures, including both $\mathbf{U}$ and $\max_{\mathbf{X}\in\mathrm{St}(n,k)}\mathrm{tr}(\mathbf{X}^\top\mathbf{A}\mathbf{X})$ that is set to $\sum_{i=1}^{k}\lambda_i$, are obtained using Matlab's EIGS function for benchmarking. For each measure, lower values indicate higher quality.

Given a solution $\tilde{\mathbf{X}}_0$ of low precision at $E(\tilde{\mathbf{X}}_0) \leq 10^{-6}$, one solver targets a solution of double precision, that is, $E(\tilde{\mathbf{X}}) \leq 10^{-12}$ or $\mathbf{\Theta}(\tilde{\mathbf{X}})/k \leq 10^{-12}$ or $1 - \cos^2\angle_{\max}(\mathbf{U},\mathbf{X}) \leq 10^{-12}$. Each algorithm terminates when the precision requirement is met or the maximum number of epoches (set as 20) is reached.

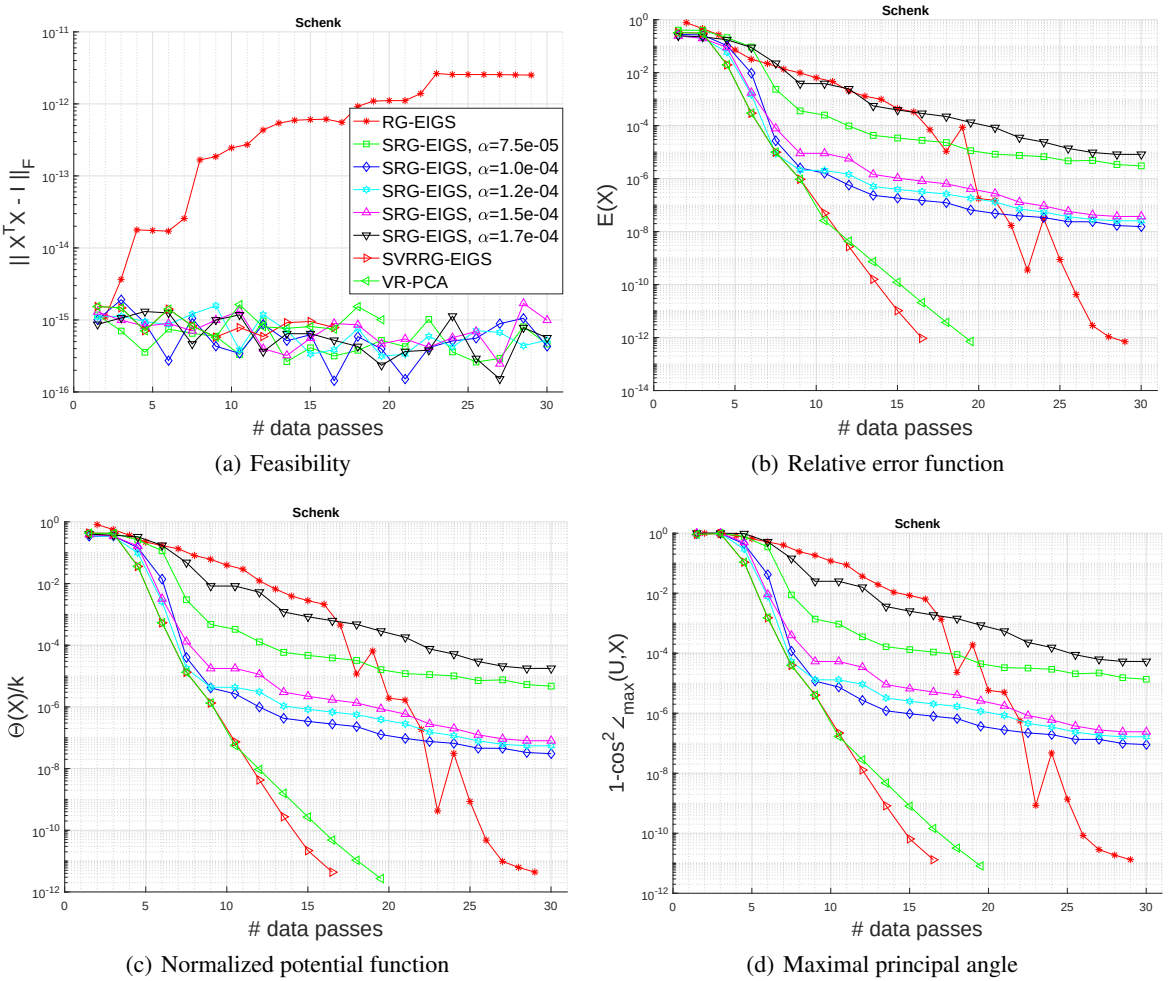
We report the convergence curves in terms of each mea-

[5] Strictly speaking, the linear rate is achievable only if $\epsilon = o(\tau^2)$.

[6] optman.blogs.rice.edu/

[7] It handles data $\mathbf{B} \in \mathbb{R}^{n\times m}$ consisting of m n-dimensional samples in $\mathbb{R}^n$, instead of data $\mathbf{A} \in \mathbb{R}^{n\times n}$ directly, where $\mathbf{A} = \mathbf{B}\mathbf{B}^\top$.

[8] www.cise.ufl.edu/research/sparse/matrices/



(a) Feasibility

(b) Relative error function

(c) Normalized potential function

(d) Maximal principal angle

Figure 1: Performance on Schenk. Note that the y-axis in each figure is in log scale.

sure, on which empirical convergence rates of the algorithms can be observed. Figure 1 reports the performance of different algorithms. In terms of the feasibility, all solvers but RG-EIGS perform well, while RG-EIGS produces much poorer results. This is because the Cayley transformation based retraction used therein relies heavily on the Sherman-Morrison-Woodbury formula, which suffers from the numerical instability. From Figure 1(b) to Figure 1(d), we observe similar convergence trends for each algorithm under the three different measures. All the four algorithms improve their solutions with more iterations. There are several exceptions in RG-EIGS. This is due to the non-monotone step size used in its implementation. We also observe that SRG-EIGS presents an exponential convergence rate at an early stage thanks to a relatively large learning rate. However, it subsequently steps into a long period of sub-exponential convergence, which leads to small progress towards the optimal solution. In contrast, VR-PCA and SVRRG-EIGS inherit the initial momentum from SRG-EIGS and keep the exponential convergence rate throughout the entire process. This enables it to approach the optimal solution at a fast speed. In particular, compared to VR-PCA, SVRRG-EIGS takes less passes over data to reach the required precision and is about one order of magnitude more accurate after the same number of data passes at a later stage. RG-EIGS has a different trend. It converges sub-exponentially at the beginning and performs worst. Though it converges fast subsquently, it still needs more passes over data than SVRRG-EIGS and VR-PCA in order to achieve the target precision.

More experimental results can be found in (Jiang et al., 2017), where experiments conducted on synthetic datsets also show the superior performance of our algorithm to that of VR-PCA when a large k is chosen.

7 Conclusion

In this paper, we proposed a fast stochastic Riemannian eigensolver by leveraging the recently proposed variance reduction technique in the stochastic Riemannian gradient optimization scheme. In addition to Riemannian gradients and retractions, the operation of vector transport as a new ingredient needs to be introduced in order to generalize SVRG properly to Riemannian manifolds. It has been deployed for the eigenvalue problem to yield a new eigensolver, i.e., the SVRRG-EIGS algorithm. And built upon the analysis of VR-PCA, we proved its local, eigengap-dependent and linear convergence at an improved rate in theory and with empirical support. However, as the learning rate is hand-tuned currently, we find it a difficult task in practice. In the future, we may conduct more empirical investigations towards automatically adjusting learning rates. And it is also important to address SVRRG-EIGS' limitations for $k \geq 1$, such as nontrivial initialization and eigen-gap dependence.

Acknowledgements

We thank anonymous reviewers for suggestions that help improve the quality of the paper. This research is supported in part by the funding from King Abdullah University of Science and Technology (KAUST), the AcRF Tier-1 Grant (RG135/14) from Ministry of Education of Singapore and the SUG Grant M4081416.020 from Nanyang Technological University.

References

P-A Absil, Robert Mahony, and Rodolphe Sepulchre. *Optimization algorithms on matrix manifolds*. Princeton University Press, 2008.

Zeyuan Allen-Zhu and Yuanzhi Li. Even faster svd decomposition yet without agonizing pain. In *Advances in Neural Information Processing Systems*, pages 974–982, 2016.

Akshay Balsubramani, Sanjoy Dasgupta, and Yoav Freund. The fast convergence of incremental pca. In *Advances in Neural Information Processing Systems*, pages 3174–3182, 2013.

Silvere Bonnabel. Stochastic gradient descent on riemannian manifolds. *IEEE Trans. Automat. Contr.*, 58(9): 2217–2229, 2013. doi: 10.1109/TAC.2013.2254619.

Petros Drineas and Michael W. Mahoney. On the nyström method for approximating a gram matrix for improved kernel-based learning. *J. Mach. Learn. Res.*, 6:2153–2175, December 2005. ISSN 1532-4435.

Alan Edelman, Tomás A. Arias, and Steven T. Smith. The geometry of algorithms with orthogonality constraints. *SIAM J. Matrix Anal. Appl.*, 20(2):303–353, April 1999. ISSN 0895-4798. doi: 10.1137/S0895479895290954.

Dan Garber, Elad Hazan, and Tengyu Ma. Online learning of eigenvectors. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, pages 560–568, 2015.

Dan Garber, Elad Hazan, Chi Jin, Sham M. Kakade, Cameron Musco, Praneeth Netrapalli, and Aaron Sidford. Faster eigenvector computation via shift-and-invert preconditioning. In *International Conference on Machine Learning*, pages 2626–2634, 2016.

Rong Ge, Chi Jin, Sham M. Kakade, Praneeth Netrapalli, and Aaron Sidford. Efficient algorithms for large-scale generalized eigenvector computation and canonical correlation analysis. In *International Conference on Machine Learning*, pages 2741–2750, 2016.

Gene H. Golub and Charles F. Van Loan. *Matrix Computations (3rd Ed.)*. Johns Hopkins University Press, Baltimore, MD, USA, 1996. ISBN 0-8018-5414-8.

N. Halko, P. G. Martinsson, and J. A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Rev.*, 53(2):217–288, May 2011. ISSN 0036-1445. doi: 10.1137/090771806.

Prateek Jain, Chi Jin, Sham M. Kakade, Praneeth Netrapalli, and Aaron Sidford. Streaming PCA: matching matrix bernstein and near-optimal finite sample guarantees for oja's algorithm. In *Proceedings of the 29th Conference on Learning Theory, COLT 2016, New York, USA, June 23-26, 2016*, pages 1147–1164, 2016. URL `http://jmlr.org/proceedings/papers/v49/jain16.html`.

Bo Jiang, Shiqian Ma, Anthony Man-Cho So, and Shuzhong Zhang. Vector transport-free svrg with general retraction for riemannian optimization: Complexity analysis and practical implementation. *arXiv preprint arXiv:1705.09059*, 2017.

Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 315–323. Curran Associates, Inc., 2013.

Hiroyuki Kasai, Hiroyuki Sato, and Bamdev Mishra. Riemannian stochastic variance reduced gradient on grassmann manifold. *CoRR*, abs/1605.07367, 2016.

John M. Lee. *Introduction to smooth manifolds*. Springer, 2012.

A. S. Lewis. Convex analysis on the hermitian matrices. *SIAM Journal on Optimization*, 6:164–177, 1996.

Michael Mitzenmacher and Eli Upfal. *Probability and Computing: Randomized Algorithms and Probabilistic Analysis*. Cambridge University Press, New York, NY, USA, 2005. ISBN 0521835402.

Andrew Y. Ng, Michael I. Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In T.G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14*, pages 849–856. MIT Press, 2002.

Erkki Oja and Juha Karhunen. On stochastic approximation of the eigenvectors and eigenvalues of the expectation of a random matrix. *Journal of mathematical analysis and applications*, 106(1):69–84, 1985.

Beresford N. Parlett. *The Symmetric Eigenvalue Problem*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1998. ISBN 0-89871-402-8.

William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery. *Numerical Recipes 3rd Edition: The Art of Scientific Computing*. Cambridge University Press, New York, NY, USA, 3 edition, 2007. ISBN 0521880688, 9780521880688.

Ohad Shamir. A stochastic PCA and SVD algorithm with an exponential convergence rate. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, pages 144–152, 2015.

Ohad Shamir. Fast stochastic algorithms for SVD and PCA: convergence properties and convexity. In *International Conference on Machine Learning*, pages 248–256, 2016.

U. Torbjorn Ringertz. Eigenvalues in optimum structural design. *Institute for Mathematics and Its Applications*, 92:135, 1997.

Chong Wang, Xi Chen, Alex J Smola, and Eric P Xing. Variance reduction for stochastic gradient optimization. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 181–189. Curran Associates, Inc., 2013.

Zaiwen Wen and Wotao Yin. A feasible method for optimization with orthogonality constraints. *Math. Program.*, 142(1-2):397–434, 2013. doi: 10.1007/s10107-012-0584-1.

Zaiwen Wen, Chao Yang, Xin Liu, and Yin Zhang. Trace-penalty minimization for large-scale eigenspace computation. Technical report, RICE UNIV HOUSTON TX DEPT OF COMPUTATIONAL AND APPLIED MATHEMATICS, 2013.

J.H. Wilkinson. *The Algebraic Eigenvalue Problem*. Monographs on numerical analysis. Clarendon Press, 1988. ISBN 9780198534181. URL `https://books.google.com.sg/books?id=5wsK1OP7UFgC`.

Zhiqiang Xu and Yiping Ke. Stochastic variance reduced riemannian eigensolver. *CoRR*, abs/1605.08233, 2016a.

Zhiqiang Xu and Yiping Ke. Effective and efficient spectral clustering on text and link data. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, CIKM '16, pages 357–366, New York, NY, USA, 2016b. ACM. ISBN 978-1-4503-4073-1. doi: 10.1145/2983323.2983708. URL `http://doi.acm.org/10.1145/2983323.2983708`.

Zhiqiang Xu, Peilin Zhao, and Jianneng Cao. Matrix eigen-decomposition via doubly stochastic riemannian optimization. In *International Conference on Machine Learning*, pages 1660–1669, 2016.

Hongyi Zhang, Sashank J. Reddi, and Suvrit Sra. Riemannian SVRG: fast stochastic optimization on riemannian manifolds. In *Advances in Neural Information Processing Systems, Barcelona, Spain*, pages 4592–4600, 2016.