On Truly Block Eigensolvers via Riemannian Optimization

Zhiqiang Xu*
zhiqiang.xu@kaust.edu.sa

Xin Gao
xin.gao@kaust.edu.sa

Computer, Electrical and Mathematical Sciences and Engineering Division
King Abdullah University of Science and Technology (KAUST)

Abstract

We study theoretical properties of block solvers for the eigenvalue problem. Despite a recent surge of interest in such eigensolver analysis, truly block solvers have received relatively less attention, in contrast to the majority of studies concentrating on vector versions and non-truly block versions that rely on the deflation strategy. In fact, truly block solvers are more widely deployed in practice by virtue of its simplicity without compromise on accuracy. However, the corresponding theoretical analysis remains inadequate for first-order solvers, as only local and k-th gap-dependent rates of convergence have been established thus far. This paper is devoted to revealing significantly better or as-yet-unknown theoretical properties of such solvers. We present a novel convergence analysis in a unified framework for three types of first-order Riemannian solvers, i.e., deterministic, vanilla stochastic, and stochastic with variance reduction, that are to find top-k eigenvectors of a real symmetric matrix, in full generality. In particular, the issue of zero gaps between eigenvalues, to the best of our knowledge for the first time, is explicitly considered for these solvers, which brings new understandings, e.g., the dependence of convergence on gaps other than the k-th one. We thus propose the concept of generalized k-th gap. Three types of solvers are proved to converge to a globally optimal solution at a global, generalized k-th gap-dependent, and linear or sub-linear rate.

*Corresponding author.

Proceedings of the 21^{st} International Conference on Artificial Intelligence and Statistics (AISTATS) 2018, Lanzarote, Spain. PMLR: Volume 84.

1 INTRODUCTION

The algebraic eigenvalue problem (Wilkinson, 1988), as one of the most fundamental problems in computational mathematics, has been studied for a century. This classical problem, in practice, is often to find a relatively small number k of top eigenvectors of a given symmetric matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ ($1 \leq k \ll n$), and can be formulated as trace maximization in the following form:

$$\max_{\mathbf{X}\in\mathbb{R}^{n\times k}:\mathbf{X}^\top\mathbf{X}=\mathbf{I}} f(\mathbf{X}) = \frac{1}{2}\mathrm{tr}(\mathbf{X}^\top\mathbf{A}\mathbf{X}). \tag{1}$$

It has found numerous applications in science and engineering computing, such as structural analysis (Torbjorn Ringertz, 1997), dynamical control systems (Helmke and Moore, 2012), combinatorial optimization (Mohar and Poljak, 1993), data mining and machine learning (Ng et al., 2002), just to name a few. When $k = 1$, a solver for Problem (1) is referred to as a vector solver that aims only at the leading eigenvector, such as the power method and Lanczos algorithm (Golub and Van Loan, 1996) from numerical algebra as well as Oja's algorithm (Oja and Karhunen, 1985) in the incremental PCA setting. When $k \geq 1$ the solver then is said to be a block solver such as simultaneous iteration, QR iteration, the block Lanczos algorithm (Golub and Van Loan, 1996), randomized singular value decomposition (Halko et al., 2011), the noisy power method (Hardt and Price, 2014), and randomized block krylov methods (Musco and Musco, 2015). Due to the nature of finding top-k eigenvectors simultaneously, these solvers are termed to be truly block. They are different from non-truly block ones that rely on the projection deflation technique (Mackey, 2009) to find eigenvectors one by one in descending order of corresponding eigenvalues. In this paper, we focus on truly block solvers, as they are more widely deployed than non-truly ones in practice, by virtue of its simplicity without compromise on accuracy. There has been a recent surge of interest in such eigensolver analysis which is centred on vector versions

(Balsubramani et al., 2013; Shamir, 2015, 2016b; Garber et al., 2016; Wang et al., 2016; Lei et al., 2016; Wang et al., 2017; Gao et al., 2017). It was either mentioned (Wang et al., 2017) or by default in these recent studies on vector solvers that their analyses hold for the block version by the deflation strategy. There also exists the work of the block solver in which the analysis is explicitly built on this strategy (Allen-Zhu and Li, 2016). Thus, they are all subsumed by the category of non-truly block solvers. For truly block ones, on one hand, it is unclear so far whether those recent analyses on vector solvers remain valid as there has been no such attempt except for Shamir (2015) and Ge et al. (2016). On the other hand, the aforementioned truly block solvers are not designed to work in the stochastic setting, albeit with nice properties, e.g., global/gap-free/optimal convergence rates (Halko et al., 2011; Musco and Musco, 2015), in their current forms. The state-of-the-art with the intersection of both worlds, i.e., stochastic truly block solvers, such as the noisy power method (Hardt and Price, 2014) and block VR-PCA (Shamir, 2016a), currently can only make it to a local/global, k-th gap-dependent, and linear rate. The noisy power method entails restrictive assumptions on noises, e.g., k-th gap-dependent and ϵ-small norm, while the block VR-PCA is local. They both are k-th gap-dependent. In addition, when the k-th gap is zero[1], the previous gap-free analyses (Musco and Musco, 2015; Allen-Zhu and Li, 2016) uniformly assert that gaps play no role in characterizing convergence. However, this would hardly be true, as globally optimal solutions are not unique any more and then nearest positive gaps might take effect instead. Therefore, natural questions that arise are if the convergence analysis of a stochastic truly block solver can be strengthened to achieve stronger results (e.g., global convergence) and if we can squeeze out any refreshing information from the k-th gap-free analysis (i.e., free from the dependence on the k-th gap). This paper is devoted to investigations on these questions and the quick answer to them is affirmative.

We present a simple yet effective theoretical analysis for Problem (1) via first-order Riemannian optimization. Three types of Riemannian gradient solvers, i.e., deterministic, vanilla stochastic, and stochastic with variance reduction, are cast into a unified framework and analyzed in full generality. Instead of the commonly used chordal distance (Shamir, 2016a; Xu et al., 2017), i.e., $\Theta(\mathbf{X}, \mathbf{V}_k) = k - \sum_{j=1}^{k} \cos^2 \theta_j(\mathbf{X}, \mathbf{V}_k)$, a novel potential function is defined by a variant of the Binet-Cauchy distance on Stiefel manifolds, i.e., $\Psi(\mathbf{X}, \mathbf{V}_l) = 1 - \prod_{j=1}^{\min\{k,l\}} \cos^2 \theta_j(\mathbf{X}, \mathbf{V}_l)$, where $\mathbf{X} \in \mathrm{St}(n, k)$ (see Section 3 for notations), $\mathbf{V}_l$ consists of top-l eigenvectors of $\mathbf{A}$, and $\theta_j(\mathbf{X}, \mathbf{V}_l)$ represents the j-th principal angle between $\mathbf{X}$ and $\mathbf{V}_l$ (see Section 4.1 for definitions). In particular, we make the following contributions:

- We propose a novel potential function for the convergence analysis of first-order truly block eigensolvers. It is key to global convergence and handling zero gaps.
- Different from the gap-free analysis, we figure out the surrogate role of the k-th gap in characterizing convergence when it is zero. To unify two distinct cases, i.e., k-th gap-dependent and -free, we propose generalized k-th gap. Particularly, larger generalized k-th gap reap faster convergence.
- A deterministic Riemannian gradient truly block eigensolver with constant/diminishing step-sizes is shown to converge to a globally optimal solution at a global, generalized k-th gap-dependent rate of type $O(\log \frac{1}{\epsilon})/O(\frac{1}{\epsilon})$.
- A stochastic Riemannian gradient truly block eigensolver is shown to converge to a globally optimal solution at a global, generalized k-th gap-dependent rate of type $O(\frac{1}{\epsilon})$.
- A RSVRG truly block eigensolver with constant/diminishing step-sizes is shown to converge to a globally optimal solution at a global, generalized k-th gap-dependent rate of type $O(\log \frac{1}{\epsilon})/O(\frac{1}{\epsilon})$.

The rest of the paper is organized as follows. Section 2 reviews the literature work with a focus on recent studies. Preliminaries about Riemannian optimization and three types of first-order truly block Riemannian solvers are presented in Section 3, which then is followed by theoretical analysis in Section 4. The paper concludes with discussions in Section 5.

2 RELATED WORK

There is a vast literature on solvers for the eigenvalue probelm due to its century long history. We concentrate only on those recent closely related studies, and briefly discuss three versions of recent solvers (i.e., vector, non-truly block, and truly block) and the gap-free analysis.

Recent eigensolver analyses mainly focus on the vector case. Balsubramani et al. (2013) gave finite-sample convergence rates of two schemes of the incremental PCA, i.e., Krasulina (1969), Oja and Karhunen (1985).

[1] Then the multiplicity of the k-th eigenvalue is greater than 1.

Table 1: Comparison of Truly Block Eigensolvers

Solver	Convergence	k-th gap	rate	solution	stoc	VR
Hardt et al. (2014)	global	dependent	$O(\log \frac{1}{\epsilon})$	globally optimal	yes	no
Musco et al. (2015)	global	dependent or free	$O(\log \frac{1}{\epsilon})$ or $\frac{1}{\epsilon^{3/2}}$	globally optimal	no	no
Shamir (2016a)	local	dependent	$O(\log \frac{1}{\epsilon})$	globally optimal	yes	yes
Ge et al. (2016)	global	dependent	$O(\log \frac{1}{\epsilon})$	globally optimal	no	no
Absil et al. (2008)	global or local	free	$O(\log \frac{1}{\epsilon})$	critical or optimal	no	no
Bonnabel (2013)	global	free	unknown	critical point	yes	no
Zhang et al. (2016)	global	free	$1/\epsilon$	critical point	yes	yes
Xu et al. (2016)	local	dependent	$1/\epsilon$	globally optimal	yes	no
Liu et al. (2016)	local	dependent	$O(\log \frac{1}{\epsilon})$	globally optimal	no	no
Xu et al. (2017)	local	dependent	$O(\log \frac{1}{\epsilon})$	globally optimal	yes	yes
Theorem 4.1	global	depedent or free	$O(\log \frac{1}{\epsilon})$ or $\frac{1}{\epsilon}$	globally optimal	no	no
Theorem 4.2	global	depedent or free	$1/\epsilon$	globally optimal	yes	no
Theorem 4.3	global	depedent or free	$O(\log \frac{1}{\epsilon})$ or $\frac{1}{\epsilon}$	globally optimal	yes	yes

(where stoc and VR stand for stochastic and variance reduction, respectively)

The rates are global, k-th gap-dependent and sub-linear $O(\frac{1}{\epsilon})$. Shamir (2015) proposed the VR-PCA which is the first stochastic variance reduced PCA, and proved its local, k-th gap-dependent and linear convergence rate $O(\log \frac{1}{\epsilon})$. Garber et al. (2016) presented a robust analysis of the shift-and-invert preconditioning method for faster eigenvector computation and achieved a global, k-th gap-dependent and linear convergence rate. Note that this is not a conventional stochastic algorithm like stochastic gradient descent (SGD). The merit of the method is that it enables the problem to be decomposed into a series of linear system problems that can leverage fast stochastic gradient methods, e.g., SVRG (Johnson and Zhang, 2013). This idea or similar ones were also exploited or extended to the generalized eigenvalue problem and canonical correlation analysis (Ge et al., 2016; Wang et al., 2016, 2017; Gao et al., 2017). The analyses above have not been extended to truly block cases except for Shamir (2015) and Ge et al. (2016), as such an extension is often non-trivial though the non-truly block case is apparent. Shamir (2016a) extended the VR-PCA to the truly block setting with the same theoretical guarantee stated, while Ge et al. (2016) analyzed the deterministic truly block CCA and achieved global, k-th gap-dependent and linear rates of convergence. Allen-Zhu and Li (2016) presented an improved analysis of the shift-and-invert method for stochastic non-truly block singular value decomposition.

There are relatively less recent studies on a truly block solver. Halko et al. (2011) studied randomized SVD which first makes use of random sampling to compress the input matrix and then does the job in the reduced space. Musco and Musco (2015) proposed a stronger and faster approximate SVD by randomized block Krylov methods and proved global, k-th gap-dependent or gap-free, and linear $O(\log \frac{1}{\epsilon})$ or sub-linear $O(\frac{1}{\sqrt{\epsilon}})$ rates. However, they don't have stochastic versions. Hardt and Price (2014) proposed the noisy power method which is truly block with a global, k-th gap-dependent and linear convergence rate. Although the stochastic version is included as a special case, the rate comes with restrictive assumptions. For example, the noise norm is required to be ϵ-small, which can hardly hold at the beginning of stochastic algorithms even with variance reduction.

We now turn to first-order truly block solvers, which is more closely related to our focus in this paper. Absil et al. (2008) provided analysis for general line-search based Riemannian first-order methods with global and linear convergence to critical points or local and linear convergence to globally optimal points. Global convergence to critical points for the SGD on Riemannian manifolds was established by Bonnabel (2013). Zhang et al. (2016) showed global and sub-linear $O(\frac{1}{\epsilon})$ convergence of the RSVRG (i.e., Riemannian SVRG) to critical points. They all aim to address general first-order Riemannian optimization, and hence are applicable to Problem (1) and gap-free. In addition, the doubly stochastic truly block Riemannian solver was shown to have a local, k-th gap-dependent and sub-linear $O(\frac{1}{\epsilon})$ rate of convergence by sampling the data and variable simultaneously Xu et al. (2016), and a linear rate was further achieved by variance reduction (Xu et al., 2017). In a distinct fashion, i.e., by showing an explicit Łojasiewicz exponent at $\frac{1}{2}$, Liu et al. (2016) established a local, k-th gap-depedent and linear rate of convergence of Riemannian line-search methods for quadratic problems with orthogonality constraints including Problem (1) as a special case. It is worth not-

ing that the block VR-PCA is subsumed by first-order truly block solvers as well. Our work falls into this category. Deterministic, vanilla stochastic, and RSVRG truly block solves are proved to converge to a globally optimal solution at global, k-th gap-dependent/-free, and linear $O(\log \frac{1}{\epsilon})$/sub-linear $O(\frac{1}{\epsilon})$ rates. A comparison of truly block solvers is summarized in Table 1.

There have been a few gap-free studies as well (Musco and Musco, 2015; Allen-Zhu and Li, 2016). Note that the gap-freeness in previous analysis means that convergence does not depend on any of $(n-1)$ gaps. Our analysis shows that the dependence on gaps is inherent in characterizing convergence of such solvers. Specifically, it depends either on the k-th gap itself or on its nearest positive ones. And they are unified in a concept called generalized k-th gap. It brings new understandings other than the gap-free analysis. While our analysis is based on first-order Riemannian optimization, the idea might be instrumental to other scenarios as well.

3 PRELIMINARIES

In this section, we present some basic background knowledge on Riemannian optimization. It is then followed by a brief introduction to three types of first-order Riemannian gradient solvers that are to be analyzed.

3.1 Riemannian Optimization

For a d-dimensional Riemmanian manifold (Lee, 2012) $\mathcal{M}$ its tangent space at a point $\mathbf{X} \in \mathcal{M}$, denoted as $T_{\mathbf{X}}\mathcal{M}$, is a d-dimensional Euclidean space $\mathbb{R}^d$ tangential to $\mathcal{M}$ at $\mathbf{X}$. The Riemannian gradient of a function $f(\mathbf{X})$ on $\mathcal{M}$, denoted as $\tilde{\nabla} f(\mathbf{X})$, 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 $\mathbf{X} \in \mathcal{M}$. $\tilde{\nabla} f(\mathbf{X}) \in T_{\mathbf{X}}\mathcal{M}$ is the unique tangent vector that satisfies

$$\langle \tilde{\nabla} f(\mathbf{X}), \xi \rangle_{\mathbf{X}} = Df(\mathbf{X})[\xi]$$

for any $\xi \in T_{\mathbf{X}}\mathcal{M}$, where $Df(\mathbf{X})[\xi]$ represents the directional derivative of $f(\mathbf{X})$ in ξ. Updates in Riemannian gradient descent (Absil et al., 2008) on $\mathcal{M}$ can be written as:

$$\mathbf{X}_{t+1} = R\left(\mathbf{X}_t, \alpha_{t+1} \tilde{\nabla} f(\mathbf{X}_t)\right),$$

where $\alpha_{t+1} > 0$ is the step-size along the gradient direction at the t-th step, 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 $\mathcal{M}$. An ideal retraction is the exponential map which, however, is computationally costly in general. Its first-order approximation is used in practice. In addition, an arithmetic operator between tangent vectors at different points cannot act as in Euclidean space until they are parallel transported to the same tangent space. Likewise, only its first-order approximation of parallel transport, known as vector transport, is used. The vector transport of a tangent vector from $\mathbf{X}$ to $\mathbf{Y}$ on $\mathcal{M}$, denoted as $\mathcal{T}_{\mathbf{X}\to\mathbf{Y}}$, is a mapping from $T_{\mathbf{X}}\mathcal{M}$ to $T_{\mathbf{Y}}\mathcal{M}$. When $\mathcal{M}$ is an embedded Riemannian sub-manifold of a Euclidean space, it can be simply defined as $\mathcal{T}_{\mathbf{X}\to\mathbf{Y}}(\xi_{\mathbf{X}}) = P_{\mathbf{Y}}(\xi_{\mathbf{X}})$ where $P_{\mathbf{Y}}(\cdot)$ represents the orthogonal projector onto $T_{\mathbf{Y}}\mathcal{M}$.

Problem (1) can be treated as trace maximization on the Stiefel manifold, i.e.,

$$\mathrm{St}(n,k) = \{\mathbf{X} \in \mathbb{R}^{n\times k} : \mathbf{X}^\top \mathbf{X} = \mathbf{I}\},$$

where $\mathbf{I}$ represents an identity matrix of appropriate size. It is an embedded Riemannian sub-manifold of $\mathbb{R}^{n\times k}$ with metric $\langle \xi, \eta \rangle_{\mathbf{X}} = \mathrm{tr}(\xi^\top \eta)$, for $\xi, \eta \in T_{\mathbf{X}}\mathrm{St}(n,k)$. The projection of $\xi \in \mathbb{R}^{n\times k}$ onto $T_{\mathbf{X}}\mathrm{St}(n,k)$ is given by

$$P_{\mathbf{X}}(\xi) = (\mathbf{I} - \mathbf{X}\mathbf{X}^\top)\xi + \mathbf{X}\mathrm{skew}(\mathbf{X}^\top \xi).$$

In particular, Riemannian gradients can be obtained via projection of ambient Euclidean gradients $\nabla f(\mathbf{X})$, i.e., $\tilde{\nabla} f(\mathbf{X}) = P_{\mathbf{X}}(\nabla f(\mathbf{X}))$. We use the retraction defined by the polar decomposition

$$R(\mathbf{X}, \xi) = (\mathbf{X} + \xi)(\mathbf{I} + \xi^\top \xi)^{-1/2},\ \xi \in T_{\mathbf{X}}\mathrm{St}(n,k).$$

3.2 First-Order Riemannian Solvers

Solver 1: deterministic One update of the deterministic Riemannian gradient solver (Absil et al., 2008) is $\mathbf{X}_{t+1} = R(\mathbf{X}_t, \alpha_{t+1} \tilde{\nabla} f(\mathbf{X}_t))$, where $\tilde{\nabla} f(\mathbf{X}_t) = (\mathbf{I} - \mathbf{X}_t \mathbf{X}_t^\top)\mathbf{A}\mathbf{X}_t$.

Solver 2: vanilla stochastic Assume that $\mathbf{A} = (1/L)\sum_{i=1}^{L} \tilde{\mathbf{A}}_i$. One update of the vanilla stochastic Riemannian gradient solver (Bonnabel, 2013) can be written as $\mathbf{X}_{t+1} = R(\mathbf{X}_t, \alpha_{t+1}\, g(\mathbf{X}_t, y_{t+1}))$, where y_{t+1} is a random variable uniformly sampled from $\{1, 2, \cdots, L\}$ at the t-th step, and $g(\mathbf{X}_t, y_{t+1})$ is stochastic Riemannian gradient defined as $g(\mathbf{X}_t, y_{t+1}) = (\mathbf{I} - \mathbf{X}_t \mathbf{X}_t^\top)\mathbf{A}_{t+1}\mathbf{X}_t$ with $\mathbf{A}_{t+1} = \tilde{\mathbf{A}}_{y_{t+1}}$ and satisfying $\mathbb{E}[g(\mathbf{X}_t, y_{t+1})|\mathbf{X}_t] = \tilde{\nabla} f(\mathbf{X}_t)$ with expectation taken with respect to y_{t+1}. Two gradients are related as follows

$$\begin{aligned} & g(\mathbf{X}_t, y_{t+1}) \\ = \ & \tilde{\nabla} f(\mathbf{X}_t) + (\mathbf{I} - \mathbf{X}_t \mathbf{X}_t^\top)(\mathbf{A}_{t+1} - \mathbf{A})\mathbf{X}_t, \end{aligned} \quad (2)$$

where the last term is stochastic and zero-mean conditioned on $\mathbf{X}_t$. In particular, the step-size sequence satisfies $\sum_t \alpha_t = +\infty$ and $\sum_t \alpha_t^2 < +\infty$ so that α_t decays to zero but not too fast. In practice, it is often set to $\alpha_t = \vartheta(t+\tau)^{-\kappa}$ with constants $\vartheta > 0$, $\tau \geq 0$ and $\kappa \in (0.5, 1]$.

Algorithm 1 RSVRG eigensolver

Input: matrix $\mathbf{A}$, initial iterate $\tilde{\mathbf{X}}_0$, step-size α, epoch length m
for $s = 1, 2, \cdots$ **do**
  $\tilde{\nabla} f(\tilde{\mathbf{X}}_{s-1}) = (\mathbf{I} - \tilde{\mathbf{X}}_{s-1}\tilde{\mathbf{X}}_{s-1}^\top)\mathbf{A}\tilde{\mathbf{X}}_{s-1}$
  $\mathbf{X}_0 = \tilde{\mathbf{X}}_{s-1}$
  for $t = 1, 2, \cdots, m$ **do**
    Pick $y_t \in \{1, 2, \ldots, L\}$ uniformly at random
    $\mathbf{X}_t = R\left(\mathbf{X}_{t-1}, \alpha\, g(\tilde{\mathbf{X}}_{s-1}, \mathbf{X}_{t-1}, y_t)\right)$
  end for
  $\tilde{\mathbf{X}}_s = \mathbf{X}_m$
end for

Solver 3: variance reduction One update of RSVRG solver (Xu et al., 2017) can be written as $\mathbf{X}_{t+1} = R(\mathbf{X}_t, \alpha_{t+1}\, g(\tilde{\mathbf{X}}, \mathbf{X}_t, y_{t+1}))$, where the RSVRG is

$$\begin{aligned} & g(\tilde{\mathbf{X}}, \mathbf{X}_t, y_{t+1}) \\ = \;& g(\mathbf{X}_t, y_{t+1}) - \mathcal{T}_{\tilde{\mathbf{X}}\to\mathbf{X}_t}((g(\tilde{\mathbf{X}}, y_{t+1}) - \tilde{\nabla} f(\tilde{\mathbf{X}}))\mathbf{B}_t). \end{aligned}$$

$\tilde{\mathbf{X}}$ is the estimated $\mathbf{X}$ at the snapshot point after every m steps. $\mathbf{B}_t = \check{\mathbf{P}}_t\hat{\mathbf{P}}_t^\top$ is obtained from the SVD: $\mathbf{X}_t^\top\tilde{\mathbf{X}} = \hat{\mathbf{P}}_t\mathbf{\Lambda}_t\check{\mathbf{P}}_t^\top$. In addition, $\mathcal{T}_{\tilde{\mathbf{X}}\to\mathbf{X}_t}$ transports gradients from tangent space $T_{\tilde{\mathbf{X}}}\mathrm{St}(n,k)$ to another $T_{\mathbf{X}_t}\mathrm{St}(n,k)$. Using Equation (2), this gradient can be expanded as follows:

$$\begin{aligned} & g(\tilde{\mathbf{X}}, \mathbf{X}_t, y_{t+1}) \\ = \;& \tilde{\nabla} f(\mathbf{X}_t) + (\mathbf{I} - \mathbf{X}_t\mathbf{X}_t^\top)(\mathbf{A}_{t+1} - \mathbf{A})\mathbf{X}_t - \\ & P_{\mathbf{X}_t}((\mathbf{I} - \tilde{\mathbf{X}}\tilde{\mathbf{X}}^\top)(\mathbf{A}_{t+1} - \mathbf{A})\tilde{\mathbf{X}}\mathbf{B}_t) \\ = \;& \tilde{\nabla} f(\mathbf{X}_t) + (\mathbf{I} - \mathbf{X}_t\mathbf{X}_t^\top)(\mathbf{A}_{t+1} - \mathbf{A})(\mathbf{X}_t - \tilde{\mathbf{X}}\mathbf{B}_t) + \\ & (\mathbf{I} - \mathbf{X}_t\mathbf{X}_t^\top)\tilde{\mathbf{X}}\tilde{\mathbf{X}}^\top(\mathbf{A}_{t+1} - \mathbf{A})\tilde{\mathbf{X}}\mathbf{B}_t - \\ & \mathbf{X}_t\mathrm{skew}(\mathbf{X}_t^\top(\mathbf{I} - \tilde{\mathbf{X}}\tilde{\mathbf{X}}^\top)(\mathbf{A}_{t+1} - \mathbf{A})\tilde{\mathbf{X}}\mathbf{B}_t), \end{aligned}$$

which is related to the full gradient $\tilde{\nabla} f(\mathbf{X}_t)$ as well, albeit in a more complicated form. Note that $g(\tilde{\mathbf{X}}, \mathbf{X}_t, y_{t+1}) - \tilde{\nabla} f(\mathbf{X}_t)$ is also stochastic and zero-mean conditioned on $\mathbf{X}_t$. Algorithmic steps are included in Algorithm 1 for reference.

4 Theoretical Analysis

Before proceeding with the analysis, necessary and important notions and notations are introduced including the potential function for measuring the closeness of the iterate $\mathbf{X}$ to a globally optimal solution. Main results are then presented and followed by proofs. Other proofs are deferred to the supplementary material.

4.1 Notions and Notations

Suppose that $\mathbf{A}$ has eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ and corresponding eigenvectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n \in \mathbb{R}^{n\times 1}$. Denote $\mathbf{\Sigma}_i = \mathrm{diag}(\lambda_1, \cdots, \lambda_i)$ and $\mathbf{V}_i = (\mathbf{v}_1, \cdots, \mathbf{v}_i)$, $i = 1, 2, \cdots, n$. Let $\mathcal{V} = \mathcal{V}_k$ be the set of globally optimal solutions to Problem (1), i.e.,

$$\mathcal{V}_k = \{\mathbf{V} \in \mathrm{St}(n,k) : \mathbf{AV} = \mathbf{V}\mathbf{\Sigma}_k\}.$$

For other $l \neq k$, $\mathcal{V}_l$ can be similarly defined and assume that $\mathcal{V}_0 = \emptyset$. For convenience, a k-dimensional subspace and one of its orthonormal bases $\mathbf{V} \in \mathrm{St}(n,k)$ are used interchangeably and the concrete meaning is clear from the context. If $\lambda_k > \lambda_{k+1}$ then $\mathbf{V} \in \mathcal{V}$ is unique, otherwise the size of $\mathcal{V}$ depends on the algebraic/geometric multiplicity of λ_k. In any case, it suffices for our purpose to find a single $\mathbf{V} \in \mathcal{V}$.

Define the l-th gap $\Delta_l = \lambda_l - \lambda_{l+1}$ for $l = 1, \cdots, n-1$ and $\Delta_0 = +\infty$, $\Delta_n = +\infty$. Further, let

$$\begin{aligned} k_{\min} &= \max\{l : \Delta_l > 0,\ l = 0, 1, \cdots, k-1\}, \\ k_{\max} &= \min\{l : \Delta_l > 0,\ l = k, \cdots, n\}. \end{aligned}$$

Then $k_{\max} - k_{\min}$ is the algebraic multiplicity of λ_k. If $\Delta_k > 0$, i.e., $k_{\max} = k$, then the analysis is termed to be k-th gap-dependent, otherwise it is k-th gap-free and $\Delta_{k_{\min}}$, $\Delta_{k_{\max}}$ are its nearest positive gaps. Note that we can always assume without loss of generality that

$$0 < \min\{\Delta_{k_{\min}}, \Delta_{k_{\max}}\} < +\infty,$$

otherwise $k_{\max} - k_{\min} = n$. $f(\mathbf{X})$ then is a constant function equal to $k\lambda_k$ and Problem (1) is trivial. We now can define generalized k-th gap as

$$\Delta_g = \begin{cases} \Delta_k, & \Delta_k > 0 \\ \min\{\Delta_{k_{\min}}, \Delta_{k_{\max}}\}, & \text{otherwise} \end{cases},$$

which is always positive. As we will see shortly, the convergence is always generalized k-th gap-dependent.

For any $\mathbf{X} \in \mathrm{St}(n,k)$, $\mathbf{Y} \in \mathrm{St}(n,l)$, since $\|\mathbf{X}^\top\mathbf{Y}\|_2 \leq \|\mathbf{X}\|_2\|\mathbf{Y}\|_2 = 1$, we can define principal angles (Golub and Van Loan) $0 \leq \theta_1 \leq \theta_2 \leq \cdots \leq \theta_{\min\{k,l\}} \leq \frac{\pi}{2}$ between them by singular values of $\mathbf{X}^\top\mathbf{Y}$, i.e., $\cos\theta_j(\mathbf{X}, \mathbf{Y}) = \sigma_j(\mathbf{X}^\top\mathbf{Y})$, $j = 1, \cdots, \min\{k,l\}$. We use a variant of the Binet-Cauchy distance as our potential function. It is defined by principal angles between $\mathbf{X}$ and $\mathbf{V}_l \in \mathcal{V}_l$ as follows

$$\Psi(\mathbf{X}, \mathbf{V}_l) = 1 - \prod_{j=1}^{\min\{k,l\}} \cos^2\theta_j(\mathbf{X}, \mathbf{V}_l),$$

which can be related to determinants:

$$\prod_{j=1}^{\min\{k,l\}} \cos^2\theta_j(\mathbf{X}, \mathbf{V}_l) = \begin{cases} \det(\mathbf{X}^\top\mathbf{V}_l\mathbf{V}_l^\top\mathbf{X}), & l \geq k \\ \det(\mathbf{V}_l^\top\mathbf{X}\mathbf{X}^\top\mathbf{V}_l), & l < k \end{cases}.$$

Note that when $l = k$ then $\Psi^{\frac{1}{2}}(\mathbf{X}, \mathbf{V}_l)$ recovers the Binet-Cauchy distance on Stiefel/Grassmann manifolds (Ham and Lee, 2008). It is easy to see that

$\Psi(\mathbf{X}, \mathbf{V}_l) \in [0, 1]$. Accordingly, we can define a variant of the Chordal distance as $\Theta(\mathbf{X}, \mathbf{V}_l) = \min\{k, l\} - \|\mathbf{X}^\top \mathbf{V}_l\|_F^2 \in [0, \min\{k, l\}]$.

4.2 Main Results

For ease of exposition, hereafter, we use ρ, η, ξ to represent positive numerical constants with possibly varying values at different places or cases. We are interested in globally ϵ-optimal solutions in terms of our potential function, i.e., $\Psi(\mathbf{X}, \mathbf{V}) < \epsilon$. For $\gamma \in (0, 1)$ and $1 \leq l < \frac{n+1}{2}$, let

$$p_l(\gamma) = 1 - \frac{\Gamma(\frac{l+1}{2})\Gamma(\frac{n-l+1}{2})}{\Gamma(\frac{1}{2})\Gamma(\frac{n+1}{2})}(\sin(\cos^{-1}(\gamma^{\frac{1}{2l}})))^{l(n-l)} {}_2F_1(\frac{n-l}{2}, \frac{1}{2}, \frac{n+1}{2}; \mathbf{I}_{l\times l} \sin^2(\cos^{-1}(\gamma^{\frac{1}{2l}}))),$$

where ${}_2F_1$ is the Gaussian hypergeometric function of matrix argument. Define $p_0(\gamma) = 0$ and

$$p(\gamma) = \begin{cases} p_k(\gamma), & \Delta_k > 0 \\ p_k(\gamma), & \Delta_k = 0 \text{ and } k_{\max} = n \\ p_{k_{\max}}(\gamma), & \Delta_k = 0 \text{ and } k_{\min} = 0 \\ p_{k_{\max}}(\gamma), & \Delta_k = 0,\ 0 < k_{\min} \text{ and } k_{\max} < n \end{cases}.$$

The information about $p(\gamma)$ that is useful to us is that $0 < p(\gamma) < 1$ and $p(\gamma) \to 0$ as $\gamma \to 0$. See Remark 2 with Lemma 4.7 in the supplementary material for details. Our main results are stated as follows, under the assumption that $k_{\max} \ll \frac{n+1}{2}$.

Theorem 4.1. *For any $\epsilon, \gamma \in (0, 1)$, with probability at least $1 - p(\gamma)$, Solver 1 is able to converge to a globally ϵ-optimal solution $\mathbf{V} \in \mathcal{V}$, at a global and k-th gap-dependent rate $O(\frac{1}{\gamma^2 \Delta_k^2} \log \frac{1}{\epsilon})$ with a constant step-size $0 < \alpha < \min\{\rho, \frac{\gamma \Delta_k}{\eta}\}$ or at a global and k-th gap-free rate $O(\frac{1}{\epsilon})$ with diminishing step-sizes $\alpha_t = \frac{\vartheta}{t+\tau}$ where $2\gamma \min\{\Delta_{k_{\min}}, \Delta_{k_{\max}}\}\vartheta > 1$ and $\vartheta < \tau \min\{\rho, \frac{\gamma \min\{\Delta_{k_{\min}}, \Delta_{k_{\max}}\}(1-\gamma)}{\eta}\}$.*

Theorem 4.2. *For any $\epsilon, \gamma, \iota \in (0, 1)$, with probability at least $1 - p(\gamma) - \iota$, Solver 2 is able to converge to a globally ϵ-optimal solution $\mathbf{V} \in \mathcal{V}$ in expectation at a global and generalized k-th gap-dependent rate $O(\frac{1}{\epsilon})$ with diminishing step-sizes $\alpha_t = \frac{\vartheta}{t+\tau}$ where $2\gamma \Delta_g \vartheta > 1$ and $\vartheta < \tau\rho$.*

Theorem 4.3. *For any $\epsilon, \gamma, \iota \in (0, 1)$, with probability at least $1 - p(\gamma) - \iota$, Solver 3 is able to converge to a globally ϵ-optimal solution $\mathbf{V} \in \mathcal{V}$ in expectation at a global and k-th gap-dependent rate $O(\frac{1}{\gamma^2 \Delta_k^2} \log \frac{1}{\epsilon})$ with a constant step-size $0 < \alpha < \min\{\rho, \frac{\gamma \Delta_k}{\xi}\}$ and the epoch length $m \geq \frac{\log 2}{2\alpha(\gamma \Delta_k - \alpha\eta)}$ or at a global and k-th gap-free rate $O(\frac{1}{\epsilon})$ with diminishing step-sizes $\alpha_{s,t} = \frac{\vartheta}{t+sm+\tau}$ where $m \geq 1$, $2\gamma \min\{\Delta_{k_{\min}}, \Delta_{k_{\max}}\}\vartheta > 1$ and $\vartheta < \tau\rho$.*

Remark 1 For the k-th gap-free convergence, the role of nearest positive gaps $\min\{\Delta_{k_{\min}}, \Delta_{k_{\max}}\}$ are covered up in theorems. In fact, the explicit rate is as follows:

$$\Psi(\mathbf{X}_{t+1}, \mathbf{V}_l) \leq \Psi(\mathbf{X}_0, \mathbf{V}_l)(\frac{\tau}{t+\tau+1})^{2\gamma\Delta_l\vartheta} + \frac{\eta}{t+\tau+1}.$$

Setting $l = k_{\min}$ and $l = k_{\max}$ shows that larger values of $\min\{\Delta_{k_{\min}}, \Delta_{k_{\max}}\}$ yield faster convergence to certain $\mathbf{V} \in \mathcal{V}$. In addition, the k-th gap-free convergence of Solver 3 comes from a direct application of Theorem 4.2 for Solver 2, as RSVRG $g(\tilde{\mathbf{X}}, \mathbf{X}_t, y_{t+1}) \in T_{\mathbf{X}_t}\mathrm{St}(n, k)$ is a special case of the vanilla stochastic Riemannian gradient, i.e., $\mathbb{E}[g(\tilde{\mathbf{X}}, \mathbf{X}_t, y_{t+1})|\mathbf{X}_t] = \tilde{\nabla} f(\mathbf{X}_t)$.

Remark 2 γ is related to the initial $\mathbf{X} \in \mathrm{St}(n, k)$. We can see that larger γ yields faster convergence. In particular, we can take $\gamma = \frac{1-\Psi(\mathbf{X}_0, \mathbf{V}_l)}{2}$. However, it will be small if $\mathbf{X}_0$ is uniformly sampled from $\mathrm{Grass}(n, k)$ at random[2], especially in the high-dimensional setting. Large values of γ can be expected from a warm start to solvers. Wart-started solvers are given an initial $\mathbf{X}_0$ close to a solution and thus having a small value of $\Psi(\mathbf{X}_0, \mathbf{V}_l)$. This explains the better performance achieved by hybrid solvers (Shamir, 2016a; Xu et al., 2017). Note that small values of γ does not affect global convergence in our analysis, while their counterparts do in local convergence analysis (Shamir, 2016a; Xu et al., 2017).

4.3 Proofs

Before proofs, we present several supporting lemmas. First, wee have the following obvious lemma which holds no matter whether $\Delta_k = 0$.

Lemma 4.4. *$\mathbf{X} \in \mathcal{V}$ if and only if $\Psi(\mathbf{X}, \mathbf{V}_{k_{\min}}) = \Psi(\mathbf{X}, \mathbf{V}_{k_{\max}}) = 0$.*

As we can assume $k_{\max} - k_{\min} < n$, there are four scenarios to be handled:

- $\Delta_k > 0$. We then have $\Delta_g = \Delta_k$ and only need to show that $\Psi(\mathbf{X}_t, \mathbf{V}) < \epsilon$ for each solver as $\mathbf{V}$ is unique.
- $\Delta_k = 0$ and $k_{\min} = 0$. We then have $\Delta_g = \Delta_{k_{\max}}$ and it suffices to show that $\Psi(\mathbf{X}_t, \mathbf{V}_{k_{\max}}) < \epsilon$, as every k-dimensional subspace of $\mathbf{V}_{k_{\max}}$ constitutes a solution $\mathbf{V}$.
- $\Delta_k = 0$ and $k_{\max} = n$. We then have $\Delta_g = \Delta_{k_{\min}}$ and it suffices to show that $\Psi(\mathbf{X}_t, \mathbf{V}_{k_{\min}}) < \epsilon$,

[2] Quotient manifold $\mathrm{Grass}(n, k) = \mathrm{St}(n, k)/\mathrm{St}(k, k)$.

as the direct sum of $\mathbf{V}_{k_{\min}}$ and every $(k-k_{\min})$-dimensional subspace of $\mathbf{V}_{k_{\min}}^{\perp}$ constitutes a solution $\mathbf{V}$ where $\mathbf{V}_{k_{\min}}^{\perp}$ represents the orthogonal complement of $\mathbf{V}_{k_{\min}}$.

- $\Delta_k = 0$, $0 < k_{\min}$ and $k_{\max} < n$. We then have $\Delta_g = \min\{\Delta_{k_{\min}}, \Delta_{k_{\max}}\}$ and need to show that both $\Psi(\mathbf{X}_t, \mathbf{V}_{k_{\min}}) < \epsilon$ and $\Psi(\mathbf{X}_t, \mathbf{V}_{k_{\max}}) < \epsilon$. Then the direct sum of $\mathbf{V}_{k_{\min}}$ and every $(k-k_{\min})$-dimensional subspace of $\mathbf{V}_{k_{\max}} - \mathbf{V}_{k_{\min}}$ constitutes a solution $\mathbf{V}$, where $\mathbf{V}_{k_{\max}} - \mathbf{V}_{k_{\min}}$ represents the space $\mathbf{V}_{k_{\max}}$ after removing its intersection with $\mathbf{V}_{k_{\min}}$.

Fortunately, they can be handled in a unified way. In what follows, assume that $\Delta_l > 0$. We now present three key lemmas.

Lemma 4.5. *If $\Psi(\mathbf{X}_t, \mathbf{V}_l) < 1-\gamma$, $0 < \alpha_{t+1} < \rho$, and $0 < \gamma < 1$, we have*

$$\begin{aligned} & \mathbb{E}[\Psi(\mathbf{X}_{t+1}, \mathbf{V}_l)|\mathbf{X}_t] \\ \leq\ & \Psi(\mathbf{X}_t, \mathbf{V}_l) - 2\alpha_{t+1}(1 - \Psi(\mathbf{X}_t, \mathbf{V}_l))E(\mathbf{X}_t) \\ & + \alpha_{t+1}^2 \eta \Psi(\mathbf{X}_t, \mathbf{V}_k) + \alpha_{t+1}^2 \xi \beta_t, \end{aligned}$$

where

$$E(\mathbf{X}_t) = \begin{cases} \mathrm{tr}(\mathbf{Q}_{l,t}^\top \mathbf{\Sigma}_l \mathbf{Q}_{l,t}) - \mathrm{tr}(\mathbf{X}_t^\top \mathbf{A} \mathbf{X}_t), & l \geq k \\ \mathrm{tr}(\mathbf{\Sigma}_l) - \mathrm{tr}(\mathbf{Q}_{l,t}^\top \mathbf{X}_t^\top \mathbf{A} \mathbf{X}_t \mathbf{Q}_{l,t}), & l < k \end{cases},$$

$$\mathbf{X}_t^\top \mathbf{V}_l = \begin{cases} \mathbf{P}_{l,t} \mathbf{\Lambda}_{l,t} \mathbf{Q}_{l,t}^\top, & l \geq k \\ \mathbf{Q}_{l,t} \mathbf{\Lambda}_{l,t} \mathbf{P}_{l,t}^\top, & l < k \end{cases}$$

represents the rank-$\min\{k,l\}$ *SVD of $\mathbf{X}_t^\top \mathbf{V}_l$, and*

$$\beta_t = \begin{cases} 0, & \textit{Solver } 1 \\ 1, & \textit{Solver } 2 \\ \Psi(\mathbf{X}_t, \mathbf{V}_k) + \Psi(\tilde{\mathbf{X}}_{s-1}, \mathbf{V}_k), & \textit{Solver } 3 \end{cases}.$$

This lemma characterizes a unified (stochastic) recurrence relation for three solvers.

Lemma 4.6.

$$\Delta_l \Psi(\mathbf{X}_t, \mathbf{V}_l) \leq E(\mathbf{X}_t) \leq \min\{k,l\}(\lambda_1 - \lambda_n)\Psi(\mathbf{X}_t, \mathbf{V}_l).$$

Lemma 4.7. *For a uniformly sampled point $\mathbf{Y} \in \mathrm{Grass}(n,l)$ with $l < \frac{n+1}{2}$ and $0 < \gamma < 1$, we have that $\Psi(\mathbf{Y}, \mathbf{V}_l) < 1-\gamma$ with probability at least $1 - p_l(\gamma)$.*

We are now ready to prove theorems.

Proof of Theorem 4.1

Proof. Note that $\beta_t = 0$ and first consider $\Delta_k > 0$. Setting $l = k$ and $\alpha_{t+1} = \alpha$ in Lemmas 4.5-4.6, we have

$$\begin{aligned} \Psi(\mathbf{X}_{t+1}, \mathbf{V}) &\leq \Psi(\mathbf{X}_t, \mathbf{V}) + \alpha^2 \eta \Psi(\mathbf{X}_t, \mathbf{V}) \\ &\quad - 2\alpha\Delta_k \Psi(\mathbf{X}_t, \mathbf{V})(1 - \Psi(\mathbf{X}_t, \mathbf{V})) \\ &\leq \Psi(\mathbf{X}_t, \mathbf{V})(1 - 2\alpha(\gamma\Delta_k - \eta\alpha)). \end{aligned}$$

When $0 < \alpha < \min\{\rho, \frac{\gamma\Delta_k}{\eta}\}$, $\rho = 2\alpha(\gamma\Delta_k - \eta\alpha) \in (0,1)$ and $\Psi(\mathbf{X}_t, \mathbf{V}) < 1-\gamma$ for $t \geq 1$. We then get

$$\begin{aligned} \Psi(\mathbf{X}_t, \mathbf{V}) &\leq (1-\rho)\Psi(\mathbf{X}_{t-1}, \mathbf{V}) \\ &\leq \cdots \\ &\leq (1-\rho)^t \Psi(\mathbf{X}_0, \mathbf{V}). \end{aligned}$$

By Lemma 4.7, $\Psi(\mathbf{X}_0, \mathbf{V}) < 1-\gamma$ with probability at least $1 - p(\gamma)$. Solving $(1-\rho)^T = \epsilon$ yields $T = \frac{1}{\rho}\log\frac{1}{\epsilon}$. Hence, when $0 < \alpha < \min\{\rho, \frac{\gamma\Delta_k}{\eta}\}$, Solver 1 returns a globally ϵ-optimal solution after $= \frac{1}{\rho}\log\frac{1}{\epsilon} = O(\frac{1}{\gamma^2\Delta_k^2}\log\frac{1}{\epsilon})$, with probability at least $1 - p(\gamma)$.

Consider $\Delta_k = 0$ and set $\alpha_t = \frac{\vartheta}{t+\tau}$. If $\Psi(\mathbf{X}_t, \mathbf{V}_l) < 1-\gamma$ and $0 < \alpha_{t+1} < \rho$ (e.g., $\vartheta < \tau\rho$ suffices), then

$$\begin{aligned} & \Psi(\mathbf{X}_{t+1}, \mathbf{V}_l) \\ \leq\ & \Psi(\mathbf{X}_t, \mathbf{V}_l) - 2\alpha_{t+1}\gamma\Delta_l \Psi(\mathbf{X}_t, \mathbf{V}_l) \\ & + \alpha_{t+1}^2 \eta \Psi(\mathbf{X}_t, \mathbf{V}_k) \\ \leq\ & (1 - 2\alpha_{t+1}\gamma\Delta_l)\Psi(\mathbf{X}_t, \mathbf{V}_l) + \eta\alpha_{t+1}^2. \end{aligned}$$

If $\vartheta < \tau \min\{\rho, \frac{\gamma\Delta_l(1-\gamma)}{\eta}\}$, we have

$$(1 - 2\alpha_{t+1}\gamma\Delta_k)(1-\gamma) + \eta\alpha_{t+1}^2 < 1-\gamma.$$

Accordingly, $\Psi(\mathbf{X}_t, \mathbf{V}_l) < 1-\gamma$ for $t \geq 1$ and thus the recursion holds for $t \geq 0$. By Lemma D.1 in Balsubramani et al. (2013), if $2\gamma\Delta_l\vartheta > 1$ then recursion can yield

$$\Psi(\mathbf{X}_{t+1}, \mathbf{V}_l) \leq \Psi(\mathbf{X}_0, \mathbf{V}_l)(\frac{\tau}{t+\tau+1})^{2\gamma\Delta_l\vartheta} + \frac{\eta}{t+\tau+1},$$

provided that $\Psi(\mathbf{X}_0, \mathbf{V}_l) < 1-\gamma$ and $\vartheta < \tau \min\{\rho, \frac{\gamma\Delta_l(1-\gamma)}{\eta}\}$. The proof completes by setting $l = k_{\min}$ and $l = k_{\max}$ and noting Lemma 4.7. □

Proof of Theorem 4.2

Proof. Note that $\beta_t = 1$ and $\alpha_t = \frac{\vartheta}{t+\tau}$ now. If $\Psi(\mathbf{X}_0, \mathbf{V}_l) < 1-\gamma$, then there exists ρ such that $\rho^2\sqrt{2T\log(1/\iota)} + \rho^2 T + \Psi(\mathbf{X}_0, \mathbf{V}_l) < 1-\gamma$ holds. By Lemma B.8, $\Psi(\mathbf{X}_t, \mathbf{V}_l) < 1-\gamma$ holds for $1 \leq t \leq T-1$ with probability at least $1-\iota$. As a result, as long as $\Psi(\mathbf{X}_t, \mathbf{V}_l) < 1-\gamma$ and $0 < \alpha_{t+1} < \rho$ (or $\vartheta < \tau\rho$), we will have the recursion

$$\begin{aligned} & \mathbb{E}[\Psi(\mathbf{X}_{t+1}, \mathbf{V}_l)|\mathbf{X}_t] \\ \leq\ & \Psi(\mathbf{X}_t, \mathbf{V}_l) - 2\alpha_{t+1}\gamma\Delta_l \Psi(\mathbf{X}_t, \mathbf{V}_l) \\ & + \alpha_{t+1}^2 \eta \Psi(\mathbf{X}_t, \mathbf{V}_k) + \alpha_{t+1}^2 \xi \\ \leq\ & (1 - 2\alpha_{t+1}\gamma\Delta_l)\Psi(\mathbf{X}_t, \mathbf{V}_l) + (\eta + \xi)\alpha_{t+1}^2, \end{aligned}$$

and then

$$\begin{aligned} & \mathbb{E}[\Psi(\mathbf{X}_{t+1}, \mathbf{V}_l)] \\ \leq\ & (1 - 2\alpha_{t+1}\gamma\Delta_l)\mathbb{E}[\Psi(\mathbf{X}_t, \mathbf{V}_l)] + \eta\alpha_{t+1}^2 \end{aligned}$$

holds for $1 \leq t \leq T-1$. If $\Psi(\mathbf{X}_0, \mathbf{V}_l) < 1-\gamma$, $\vartheta < \tau\rho$ and further $2\gamma\Delta_k\vartheta > 1$, as with the case of $\Delta_k = 0$ in the proof of Theorem 4.1, we have

$$\mathbb{E}[\Psi(\mathbf{X}_T, \mathbf{V}_l)] \leq \Psi(\mathbf{X}_0, \mathbf{V}_l)(\frac{\tau}{T+\tau})^{2\gamma\Delta_l\vartheta} + \frac{\eta}{T+\tau}$$

with probability at least $1-\iota$. The proof completes. □

Proof of Theorem 4.3

Proof. Note that $\beta_t = \Psi(\mathbf{X}_t, \mathbf{V}_k) + \Psi(\tilde{\mathbf{X}}_{s-1}, \mathbf{V}_k)$ now. First consider $\Delta_k > 0$. Setting $l = k$, $\alpha_{t+1} = \alpha$ and by Lemmas 4.5-4.6, we have the following recursion

$$\begin{aligned}
& \mathbb{E}[\Psi(\mathbf{X}_{t+1}, \mathbf{V})] \\
\leq\ & \mathbb{E}[\Psi(\mathbf{X}_t, \mathbf{V})] - 2\alpha\gamma\Delta_k\mathbb{E}[\Psi(\mathbf{X}_t, \mathbf{V})] \\
& +\alpha^2\eta\mathbb{E}[\Psi(\mathbf{X}_t, \mathbf{V})] \\
& +\alpha^2\xi\mathbb{E}[\Psi(\mathbf{X}_t, \mathbf{V})] + \alpha^2\xi\mathbb{E}[\Psi(\tilde{\mathbf{X}}_{s-1}, \mathbf{V})] \\
\leq\ & (1 - 2\alpha(\gamma\Delta_k - (\eta+\xi)\alpha))\mathbb{E}[\Psi(\mathbf{X}_t, \mathbf{V})] \\
& +\alpha^2\xi\mathbb{E}[\Psi(\tilde{\mathbf{X}}_{s-1}, \mathbf{V})].
\end{aligned}$$

Similarly to the proof of Theorem 4.2, if $\Psi(\tilde{\mathbf{X}}_{s-1}, \mathbf{V}) < 1-\gamma$ and $0 < \alpha < \rho$, then $\Psi(\mathbf{X}_t, \mathbf{V}) < 1-\gamma$ holds for $t = 1, \cdots, m-1$ with probability at least $1-\iota$. The recursion then holds for $t = 0, 1, \cdots, m-1$. Letting $\nu = 2\alpha(\gamma\Delta_k - \eta\alpha)$ and $\mu = \alpha^2\xi$, we can write

$$\begin{aligned}
& \mathbb{E}[\Psi(\tilde{\mathbf{X}}_s, \mathbf{V})] = \mathbb{E}[\Psi(\mathbf{X}_m, \mathbf{V})] \\
\leq\ & (1-\nu)\mathbb{E}[\Psi(\mathbf{X}_{m-1}, \mathbf{V})] + \mu\mathbb{E}[\Psi(\tilde{\mathbf{X}}_{s-1}, \mathbf{V})] \\
\leq\ & (1-\nu)^m\mathbb{E}[\Psi(\mathbf{X}_0, \mathbf{V})] \\
& +\mu\sum_{i=0}^{m-1}(1-\nu)^i\mathbb{E}[\Psi(\tilde{\mathbf{X}}_{s-1}, \mathbf{V})] \\
=\ & ((1-\nu)^m + \mu\sum_{i=0}^{m-1}(1-\nu)^i)\mathbb{E}[\Psi(\tilde{\mathbf{X}}_{s-1}, \mathbf{V})] \\
\leq\ & ((1-\nu)^m + \mu\sum_{i=0}^{\infty}(1-\nu)^i)\mathbb{E}[\Psi(\tilde{\mathbf{X}}_{s-1}, \mathbf{V})] \\
=\ & ((1-\nu)^m + \frac{\mu}{\nu})\mathbb{E}[\Psi(\tilde{\mathbf{X}}_{s-1}, \mathbf{V})],
\end{aligned}$$

where

$$(1-\nu)^m = \exp\{m\log(1-\nu)\} \leq \exp\{-m\alpha\nu\}.$$

For any $\delta \in (0, \frac{1}{2})$, letting $\mu \leq \delta\nu$ and $\exp\{-m\alpha\nu\} \leq 1-2\delta$, namely

$$\alpha \leq \frac{\gamma\Delta_k}{\xi}, \quad m \geq \frac{-\log(1-2\delta)}{\nu},$$

we can arrive at

$$\begin{aligned}
\mathbb{E}[\Psi(\tilde{\mathbf{X}}_s, \mathbf{V})] &\leq ((1-\nu)^m + \frac{\mu}{\nu})\mathbb{E}[\Psi(\tilde{\mathbf{X}}_{s-1}, \mathbf{V})] \\
&\leq (1-\delta)\mathbb{E}[\Psi(\tilde{\mathbf{X}}_{s-1}, \mathbf{V})].
\end{aligned}$$

Choose $\delta = \frac{1}{4}$. By recursion over s, we get that for any $\iota \in (0,1)$ and $\iota' \in (0, \frac{\iota}{S})$ it holds that

$$\mathbb{E}[\Psi(\tilde{\mathbf{X}}_S, \mathbf{V})] \leq 2^{-S}\mathbb{E}[\Psi(\tilde{\mathbf{X}}_0, \mathbf{V})]$$

with probability at least $1 - S\iota' = 1-\iota$. Solving $2^{-S} = \epsilon$ yields $S = O(\log\frac{1}{\epsilon})$ and thus $T = mS = O(\frac{1}{\gamma^2\Delta_k^2}\log\frac{1}{\epsilon})$. Noting Lemma 4.7, the k-th gap-dependent result holds.

Consider $\Delta_k = 0$ now. Similarly, we have

$$\begin{aligned}
& \mathbb{E}[\Psi(\mathbf{X}_{t+1}, \mathbf{V}_l)] \\
\leq\ & \mathbb{E}[\Psi(\mathbf{X}_t, \mathbf{V}_l)] - 2\alpha_{t+1}\gamma\Delta_l\mathbb{E}[\Psi(\mathbf{X}_t, \mathbf{V}_l)] \\
& +\alpha_{t+1}^2\eta\mathbb{E}[\Psi(\mathbf{X}_t, \mathbf{V}_k)] \\
& +\alpha_{t+1}^2\xi\mathbb{E}[\Psi(\mathbf{X}_t, \mathbf{V}_k)] + \alpha_{t+1}^2\xi\mathbb{E}[\Psi(\tilde{\mathbf{X}}_{s-1}, \mathbf{V}_k)] \\
\leq\ & (1 - 2\alpha_{t+1}\gamma\Delta_l)\mathbb{E}[\Psi(\mathbf{X}_t, \mathbf{V}_l)] + \alpha_{t+1}^2(\eta+\xi).
\end{aligned}$$

Unfolding the epoch over s and using the step-size $\alpha_{s,t} = \frac{\vartheta}{t+sm+\tau}$, Solver 2 with RSVRG is recovered. The proof of the k-th gap-free result completes. □

5 Discussions

In this paper, we studied generalized k-th gap-dependent convergence properties of three types of first-order truly block Riemannian eigensolvers. For first-order truly block solvers, this is the first time that global convergence rates are established, though they exist already with other solvers. Our analysis demonstrates that the dependence on gaps is inherent in characterizing eigensolvers' convergence behaviors. They are related to either the k-th gap itself or its nearest positive ones corresponding to linear or sublinear rates. This work can be improved in several ways. First, there are some limitations in the current analysis. For example, Lemma 4.7 requires $l < \frac{n+1}{2}$. Although in general we have $k \ll \frac{n+1}{2}$ for real problems, the case that $k_{\max} \ll \frac{n+1}{2}$ does not hold always in theory. Second, rates are not optimal. It would be a good direction to derive optimal rates matching those in other settings, e.g., non-truly solvers. Third, it is worth exploring ways of improving the quadratic dependence on the generalized k-th gap.

Acknowledgements

We would like to thank Renaud-Alexandre Pitaval for his insightful suggestion on the potential function. This research is supported by the funding from King Abdullah University of Science and Technology (KAUST).

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.

Chao Gao, Dan Garber, Nathan Srebro, Jialei Wang, and Weiran Wang. Stochastic canonical correlation analysis. *CoRR*, abs/1702.06533, 2017. URL `http://arxiv.org/abs/1702.06533`.

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.

G.H. Golub and C.F. Van Loan. *Matrix Computations*. Johns Hopkins Studies in the Mathematical Sciences. ISBN 9781421407944.

Nathan Halko, Per-Gunnar Martinsson, and Joel A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, 53(2):217–288, 2011. doi: 10.1137/090771806.

Jihun Ham and Daniel D. Lee. Grassmann discriminant analysis: a unifying view on subspace-based learning. In *International Conference on Machine Learning*, pages 376–383, 2008.

Moritz Hardt and Eric Price. The noisy power method: A meta algorithm with applications. In *Advances in Neural Information Processing Systems*, pages 2861–2869, 2014.

Moritz Hardt, Eric Price, and test. The noisy power method: A meta algorithm with applications. In *Advances in Neural Information Processing Systems*, pages 2861–2869, 2014.

Uwe Helmke and John B Moore. *Optimization and dynamical systems*. Springer Science & Business Media, 2012.

Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States.*, pages 315–323, 2013.

TP Krasulina. The method of stochastic approximation for the determination of the least eigenvalue of a symmetrical matrix. *USSR Computational Mathematics and Mathematical Physics*, 9(6):189–195, 1969.

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

Qi Lei, Kai Zhong, and Inderjit S. Dhillon. Coordinate-wise power method. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 2056–2064, 2016.

Huikang Liu, Weijie Wu, and Anthony Man-Cho So. Quadratic optimization with orthogonality constraints: Explicit lojasiewicz exponent and linear convergence of line-search methods. In *ICML*, pages 1158–1167, 2016.

Lester W Mackey. Deflation methods for sparse pca. In *Advances in neural information processing systems*, pages 1017–1024, 2009.

Bojan Mohar and Svatopluk Poljak. Eigenvalues in combinatorial optimization. In *Combinatorial and graph-theoretical problems in linear algebra*, pages 107–151. Springer New York, 1993.

Cameron Musco and Christopher Musco. Randomized block krylov methods for stronger and faster approximate singular value decomposition. In *NIPS*, pages 1396–1404, 2015.

Cameron Musco, Christopher Musco, and test. Randomized block krylov methods for stronger and faster approximate singular value decomposition. In *NIPS*, pages 1396–1404, 2015.

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.

Ohad Shamir. A stochastic PCA and SVD algorithm with an exponential convergence rate. In *International Conference on Machine Learning*, 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, 2016a.

Ohad Shamir. Convergence of stochastic gradient descent for PCA. In *International Conference on Machine Learning*, pages 257–265, 2016b.

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

Jialei Wang, Weiran Wang, Dan Garber, and Nathan Srebro. Efficient coordinate-wise leading eigenvector computation. *CoRR*, abs/1702.07834, 2017. URL `http://arxiv.org/abs/1702.07834`.

Weiran Wang, Jialei Wang, Dan Garber, and Nati Srebro. Efficient globally convergent stochastic optimization for canonical correlation analysis. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 766–774, 2016.

J. H. Wilkinson, editor. *The Algebraic Eigenvalue Problem*. Oxford University Press, Inc., New York, NY, USA, 1988. ISBN 0-198-53418-3.

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

Zhiqiang Xu, Yiping Ke, and Xin Gao. A fast stochastic riemannian eigensolver. In *UAI*, 2017.

Hongyi Zhang, Sashank J. Reddi, and Suvrit Sra. Riemannian SVRG: fast stochastic optimization on riemannian manifolds. In *NIPS*, pages 4592–4600, 2016.