Journal of Machine Learning Research 22 (2021) 1-46 Submitted 1/20; Published 4/21

On the Riemannian Search for Eigenvector Computation

Zhiqiang Xu XUZHIQIANG04@BAIDU.COM

Ping Li LIPING11@BAIDU.COM

Cognitive Computing Lab
Baidu Research
10900 NE 8th St. Bellevue, WA 98004, USA

Editor: Genevera Allen

Abstract

Eigenvector computation is central to numerical algebra and often critical to many data analysis tasks nowadays. Most research on this problem has been focusing on projection methods like power iterations, such that this category of algorithms can achieve both optimal convergence rates and cheap per-iteration costs. In contrast, search methods belonging to another main category are less understood in this respect. In this work, we consider the leading eigenvector computation as a non-convex optimization problem on the (generalized) Stiefel manifold and covers the cases for both standard and generalized eigenvectors. It is shown that the inexact Riemannian gradient method induced by the shift-and-invert preconditioning is guaranteed to converge to one of the ground-truth eigenvectors at an optimal rate, e.g., $O(\sqrt{\kappa_{\mathbf{B}} \frac{\lambda_1}{\lambda_1 - \lambda_{p+1}}} \log \frac{1}{\epsilon})$ for a pair of real symmetric matrices $(\mathbf{A}, \mathbf{B})$ with $\mathbf{B}$ being positive definite, where λ_i represents the i-th largest generalized eigenvalue of the matrix pair, p is the multiplicity of λ_1, and $\kappa_{\mathbf{B}}$ stands for the condition number of $\mathbf{B}$. The standard eigenvector computation is recovered by setting $\mathbf{B}$ to an identity matrix. Our analysis reduces the dependence on the eigengap, making it the first Riemannian eigensolver that achieves the optimal rate. Experiments demonstrate that the proposed search method is able to deliver significantly better performance than projection methods by taking advantages of step-size schemes.

Keywords: Eigenvector Computation, Generalized Eigenvalue Problem, Riemannian Optimization, Shift-and-invert Preconditioning, Optimal Convergence Rate

1. Introduction

Eigenvector computation is a fundamental problem in numerical algebra and often of central importance to a variety of scientific and engineering computing tasks 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 (Fan et al., 2018; Ng et al., 2001; Hastie et al., 2015; Hotelling, 1936). There are two main categories of methods for this problem: projection and search. Most of recent research has been focusing on projection methods, such as power iterations (Golub and Van Loan, 2013), for both faster convergence rates and cheaper iteration costs. Although Lanczos algorithms possess the optimal convergence rate $O(\sqrt{\frac{\lambda_1}{\lambda_1 - \lambda_2}} \log \frac{1}{\epsilon})$ (Parlett, 1998), it is not amenable to stochastic optimization (Xu et al., 2018a). People thus tend to develop fast

stochastic algorithms on top of power methods (Arora et al., 2013; Hardt and Price, 2014; Shamir, 2015; Garber and Hazan, 2015; Garber et al., 2016; Lei et al., 2016; Wang et al., 2018). Along this line of research, in particular, the shift-and-invert preconditioning as a classic acceleration technique has been revived recently (Garber and Hazan, 2015; Garber et al., 2016; Allen-Zhu and Li, 2016; Wang et al., 2016; Allen-Zhu and Li, 2017; Wang et al., 2018). This is because power iterations with the preconditioning have a provable reduction to approximately solving a sequence of linear systems of equations that can leverage fast and/or stochastic least-squares solvers, such as accelerated gradient descent (AGD) (Nesterov, 2004), (accelerated) randomized coordinate descent (Nesterov, 2012; Nesterov and Stich, 2017), and (accelerated) stochastic variance reduced gradient (SVRG) (Johnson and Zhang, 2013; Lin et al., 2015). This way has power methods accelerated to reach an optimal rate with a cheap iteration cost (Wang et al., 2018). In contrast, there have been no search methods that achieve optimal rates. The current rates of search methods exhibit a quadratic (Shamir, 2015; Xu et al., 2018b) or linear (Xu and Li, 2021b) dependence on the eigengap, e.g., $\frac{1}{(\lambda_1-\lambda_2)^2}$ or $\frac{1}{\lambda_1-\lambda_2}$, thus far. In this work, we try to reduce such dependence by taking a novel Riemannian optimization view with the preconditioning, covering both standard and generalized eigenvector computations. Taking the standard eigenvector computation as an example, the Riemannian optimization problem can be written as:

$$\max_{\mathbf{x}\in\mathbb{R}^{n\times 1}:\|\mathbf{x}\|_2=1} \frac{1}{2}\mathbf{x}^\top\mathbf{A}\mathbf{x}, \tag{1}$$

where $\mathbf{A}\in\mathbb{R}^{n\times n}$ is the given real symmetric matrix and the constraint constitutes a sphere or Stiefel manifold. Considering the shift-and-invert preconditioning for scaling up the relative eigengap of the matrix to be processed, $\mathbf{A}$ is replaced by $\mathbf{C}^{-1}=(\sigma\mathbf{I}-\mathbf{A})^{-1}$ above, arriving at the following problem:

$$\max_{\mathbf{x}\in\mathbb{R}^{n\times 1}:\|\mathbf{x}\|_2=1} \frac{1}{2}\mathbf{x}^\top\mathbf{C}^{-1}\mathbf{x}, \tag{2}$$

where σ is the shift parameter satisfying $\sigma>\lambda_1$ and λ_i represents the i-th largest eigenvalue of $\mathbf{A}$. Accordingly, the preconditioning gives rise to inverse-matrix-vector multiplications $\mathbf{C}^{-1}\mathbf{x}$ that are exact solutions to least-squares sub-problems in Riemannian gradients. Instead of solving sub-problems accurately for computing the multiplications, approximate solutions suffice and result in inexact Riemannian gradients. We take the form of Problem (2) for implementation, while the following equivalent form is considered for analysis:

$$\max_{\mathbf{y}\in\mathbb{R}^{n\times 1}:\mathbf{y}^\top\mathbf{C}\mathbf{y}=1} \frac{1}{2}\|\mathbf{y}\|_2^2, \tag{3}$$

where the constraint constitutes a generalized Stiefel manifold. Due to the preconditioning, the main problem are guaranteed to take $O(\log\frac{1}{\epsilon})$ iterations to converge. If the accelerated gradient descent is chosen as a subproblem solver, the per-iteration cost of the main problem, i.e., the number of iterations that is sufficient for solving subproblems, will be $O(\sqrt{\frac{\lambda_1}{\lambda_1-\lambda_{p+1}}})$, where p is the multiplicity of λ_1. The overall rate is $O(\sqrt{\frac{\lambda_1}{\lambda_1-\lambda_{p+1}}}\log\frac{1}{\epsilon})$ (see Theorem 2 and its proof). The success of the algorithm depends on an appropriate shift parameter σ which needs to be an upper bound on λ_1 and as close as possible to λ_1. Theoretically, this can

be guaranteed by the procedure introduced in Garber and Hazan (2015). The cost of this procedure is independent of the final accuracy, making the total cost dominated by the subsequent stage. Despite the theoretical guarantee on σ, the procedure is not easy to implement due to many parameters to be tuned. To this end, we follow Zhou et al. (2006) that uses a small number of steps of the Lanczos[1] algorithm to yield a proper upper bound on λ_1.

We further extend the analysis to the problem of generalized eigenvector computation for a pair of real symmetric matrices $(\mathbf{A}, \mathbf{B})$ with $\mathbf{B}$ being positive definite:

$$\max_{\mathbf{x}\in\mathbb{R}^{n\times 1}:\mathbf{x}^\top\mathbf{B}\mathbf{x}=1} \frac{1}{2}\mathbf{x}^\top\mathbf{A}\mathbf{x}. \tag{4}$$

Applying the preconditioning, we have the following problem for implementation:

$$\max_{\mathbf{x}\in\mathbb{R}^{n\times 1}:\mathbf{x}^\top\mathbf{B}\mathbf{x}=1} \frac{1}{2}\mathbf{x}^\top\mathbf{B}(\sigma\mathbf{B}-\mathbf{A})^{-1}\mathbf{B}\mathbf{x}. \tag{5}$$

The equivalent form for analysis is as follows:

$$\max_{\mathbf{y}\in\mathbb{R}^{n\times 1}:\mathbf{y}^\top(\sigma\mathbf{B}-\mathbf{A})\mathbf{y}=1} \frac{1}{2}\mathbf{y}^\top\mathbf{B}\mathbf{y}. \tag{6}$$

The overall rate is $O(\sqrt{\kappa_{\mathbf{B}}\frac{\lambda_1}{\lambda_1-\lambda_{p+1}}}\log\frac{1}{\epsilon})$, where λ_i now is the i-th largest generalized eigenvalue of the matrix pair $(\mathbf{A}, \mathbf{B})$ and $\kappa_{\mathbf{B}}$ represents the condition number of $\mathbf{B}$.

Despite the same optimal rate of the proposed search methods as recently proposed projection methods, search methods have additional privileges of getting the most of its step-size schemes. Particularly, we equip our methods with the popular Barzilai-Borwein (BB) step-size (Barzilai and Borwein, 1988; Iannazzo and Porcelli, 2017). Experiments show that the resulting search methods have significantly better performance compared to the projection methods.

The rest of the paper is organized as follows. Section 2 discusses recent literature works. Section 3 gives a brief introduction to Riemannian geometry and optimization. Section 4 presents the proposed Riemannian algorithm for the standard eigenvector computation and the analysis. It is then followed by the extension to the generalized eigenvector computation in Section 5. Experimental studies are reported in Section 6. The paper is concluded in Section 7.

2. Related Work

There is a vast literature on eigenvector computation (Wilkinson, 1988; Parlett, 1998; Saad, 2011). We focus on recent research.

Shift-and-invert preconditioning has attracted resurgent interests due to its ability to gain faster convergence rates in combination with state-of-the-art first-order optimization methods. Garber and Hazan (2015); Garber et al. (2016) initiated this line of research. They presented a robust analysis of the shift-and-invert preconditioned power method for principal component analysis (PCA) and achieved convergence rates of optimal type. However,

1. A small number of Lanczos steps do not hurt the overall performance.

the logarithmic dependence on parameters including the accuracy parameter is not good enough. Allen-Zhu and Li (2016) extended their work to the block setting by deflation via a careful analysis and considered the general case, i.e., standard eigenvector computation, while Wang et al. (2018) improved their analysis by removing an extra $\log \frac{1}{\epsilon}$ factor and advocated coordinate descent as the solver for subproblems to handle the general matrix without covariance structure. Wang et al. (2016) extended the work of Garber and Hazan (2015); Garber et al. (2016) to canonical correlation analysis (CCA). However, the analysis inherits the poly-logarithmic dependence on accuracy. Allen-Zhu and Li (2017) similarly extended the work of Garber and Hazan (2015); Garber et al. (2016) to the generalized eigenvector computation and CCA as well as their block setting. All these works follow the original framework of Garber and Hazan (2015); Garber et al. (2016). Our work deviates from this and takes a novel view of the Riemannian optimization, albeit with the same preconditioning. As we will see, it gives rise to a novel analysis accordingly.

SGD based methods Arora et al. (2013) considered the stochastic power method for PCA with convex relaxation and obtained the rate $O(\frac{1}{\epsilon^2})$. Balsubramani et al. (2013) achieved the rate $O(\frac{1}{(\lambda_1-\lambda_2)^2\epsilon})$ for PCA via the martingale analysis. Shamir (2015, 2016a) proposed the VR-PCA algorithm which extended the projected stochastic variance reduced gradient (SVRG) to the non-convex PCA problem. It has global convergence rate $O(\frac{1}{(\lambda_1-\lambda_2)^2}\log\frac{1}{\epsilon})$. Shamir (2016b) studied SGD for PCA and established its sub-linear convergence rates, $O(\frac{1}{(\lambda_1-\lambda_2)\epsilon})$ and $O(\frac{1}{\epsilon^2})$, for gap-dependent and gap-free cases, respectively.

Riemannian algorithms The constrained optimization for our problems becomes unconstrained in the Riemannian setting. Thus, many Riemannian optimization methods apply to our problems. Absil et al. (2008) provided analysis for general line-search based Riemannian first-order methods that converge globally and linearly to critical points or locally and linearly to minimizers. Wen and Yin (2013) proposed a practical curvilinear search method for the first-order optimization on Stiefel manifolds. It is characterized by a new retraction based on the Cayley-transform. Bonnabel (2013) established global convergence of Riemannian SGD to critical points. Zhang et al. (2016) proposed Riemannian SVRG and proved its global and sub-linear convergence rate $O(\frac{1}{\epsilon})$ to critical points. Specifically, Xu et al. (2017) proposed Riemannian SVRG eigensolver that converges at a local and linear rate $O(\frac{1}{(\lambda_1-\lambda_2)^2}\log\frac{1}{\epsilon})$. In a distinct fashion, i.e., by showing an explicit Łojasiewicz exponent at $\frac{1}{2}$, Liu et al. (2016) established a local and linear convergence of Riemannian line-search methods for quadratic problems defined on Stiefel manifolds, including Problem (2) as a special case. However, the explicit rate is unclear. All the rates here are not accelerated. Accelerated Riemannian gradient proposed recently (Zhang and Sra, 2018; Ahn and Sra, 2020) only supports geodesically convex problems in theory currently. Since our problems are geodesically non-convex, we opt for the preconditioning to accelerate Riemannian eigensolvers in this work.

Other algorithms Halko et al. (2011) surveyed and extended randomized algorithms for truncated singular value decomposition (SVD). Randomized singular value decomposition (RSVD) makes use of random sampling to compress the input matrix and then does the job in the reduced space. Let $\mathbf{A}$ be an $n \times d$ rectangular matrix. Musco and Musco (2015) proved the global rates, $O(\sqrt{\frac{\lambda_1}{\lambda_1-\lambda_2}}\log\frac{1}{\epsilon})$ and $\tilde{O}(\frac{\text{nnz}(\mathbf{A})}{\sqrt{\epsilon}} + \frac{n}{\epsilon} + \frac{1}{\epsilon^{3/2}})$, of the randomized

block Krylov methods (RBK) for an approximate SVD in the gap-dependent and gap-free cases, respectively. When $\frac{1}{\epsilon} > \text{nnz}(\mathbf{A})$, the rate is $O(\frac{1}{\epsilon^{3/2}})$. Moreover, when $\lambda_1 = \lambda_2$, the true convergence is not necessarily sub-linear as is indicated by the gap-free rate. Hardt and Price (2014) provided a new robust convergence analysis for the power method under the noise setting (NPM). It was shown to have the convergence rate $O(\frac{\lambda_1}{\lambda_1-\lambda_2}\log\frac{1}{\epsilon})$ with high probability. Balcan et al. (2016) extended this method to achieve an improved gap dependency by using subspace iterates of larger dimensions. Xu et al. (2018a) proposed an accelerated (stochastic) power method which uses the scaled Chebyshev polynomial to directly accelerate and achieve convergence rates of optimal type, i.e., $O(\frac{1}{\sqrt{\lambda_1-\lambda_2}}\log\frac{1}{\epsilon})$ in the deterministic setting (PM+M) and $O(\frac{1}{\sqrt{\lambda_1-\lambda_2}}\log\frac{1}{\epsilon}\log^2\frac{1}{\delta})$ with probability $1-\delta\log\frac{1}{\epsilon}$ in the PCA setting (VR-PM+M), provided that the momentum parameter β is optimal, i.e., $\beta = \frac{\lambda_2^2}{4}$. In addition, Ge et al. (2016) presented the first provable efficient algorithm (GenELin) for generalized eigenvector computation via the inexact power method, and achieved a global and linear rate $\tilde{O}(\frac{\lambda_1}{\lambda_1-\lambda_2}\sqrt{\kappa_{\mathbf{B}}}\log\frac{1}{\epsilon})$. The algorithm, applied to CCA (CCALin), is equivalent to alternating least-squares (CCA-ALS) proposed in Wang et al. (2016) for the top-1 case where the convergence analysis for CCA achieves a global and sub-linear rate $\tilde{O}((\frac{\lambda_1}{\lambda_1-\lambda_2})^2\sqrt{\kappa_{\mathbf{B}}}\log^2\frac{1}{\epsilon})$. Alternating least-squares for CCA was further studied in Xu and Li (2019, 2021a). Although the convergence rate was improved to be linear, it has an additional logarithmic factor on the spectral gap.

3. Riemannian Geometry and Optimization

Let $\mathcal{M}$ be a Riemmanian manifold (Lee, 2012) of dimension d and $T_{\mathbf{x}}\mathcal{M}$ be its tangent space at $\mathbf{x} \in \mathcal{M}$ which is a d-dimensional Euclidean space $\mathbb{R}^d$ tangential to $\mathcal{M}$ at $\mathbf{x}$. $\mathcal{M}$ is often associated with certain Riemannian metric which is a family of smoothly varying inner products on tangent spaces, i.e., $\langle\xi,\eta\rangle_{\mathbf{x}}$, where tangent vectors $\xi,\eta \in T_{\mathbf{x}}\mathcal{M}$ for any $\mathbf{x} \in \mathcal{M}$. Riemannian gradient of a function $f(\mathbf{x})$ on $\mathcal{M}$ is the unique tangent vector, i.e., $\tilde{\nabla} f(\mathbf{x}) \in T_{\mathbf{x}}\mathcal{M}$, that satisfies

$$\langle\tilde{\nabla} f(\mathbf{x}),\xi\rangle_{\mathbf{x}} = Df(\mathbf{x})[\xi] \tag{7}$$

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

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

where $\alpha_t > 0$ is the step-size at the current 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 $\mathcal{M}$. Instead of using computationally costly exponential map, cheap retractions are used, e.g., first-order approximation of exponential map. In addition, tangent vectors at different points need to be parallel transported to the same tangent space before arithmetic operations between them. However, in certain circumstances for efficiency, this operation is often omitted without sacrificing accuracy much.

For the generalized Stiefel manifold

$$\text{St}_{\mathbf{B}}(n,1) = \{\mathbf{x} \in \mathbb{R}^{n\times 1} : \mathbf{x}^\top\mathbf{B}\mathbf{x} = 1\},$$

we use the Riemannian metric

$$\langle \xi, \eta \rangle_{\mathbf{B}} = \xi^\top \mathbf{B} \eta,$$

for $\xi, \eta \in T_{\mathbf{x}}\mathrm{St}_{\mathbf{B}}(n,1)$. By the definition (7), $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top \mathbf{A}\mathbf{x}$ on $\mathrm{St}_{\mathbf{B}}(n,1)$ has the following Riemannian gradient

$$\tilde{\nabla} f(\mathbf{x}) = (\mathbf{B}^{-1} - \mathbf{x}\mathbf{x}^\top)\mathbf{A}\mathbf{x}.$$

In addition, we use the retraction defined by the generalized polar decomposition

$$R(\mathbf{x}, \xi) = \frac{\mathbf{x} + \xi}{\|\mathbf{x} + \xi\|_{\mathbf{B}}}$$

for tangent vector $\xi \in T_{\mathbf{x}}\mathrm{St}_{\mathbf{B}}(n,1)$, where $\|\mathbf{x}\|_{\mathbf{B}} = \sqrt{\mathbf{x}^\top \mathbf{B}\mathbf{x}}$. For the standard case, it suffices to set $\mathbf{B} = \mathbf{I}$.

4. Standard Eigenvector Computation

In this section, we present our shift-and-invert preconditioned Riemannian eigensolver for the standard eigenvector computation, i.e., Problem (1).

4.1 Algorithm

Given a real symmetric matrix $\mathbf{A}$, i.e., $\mathbf{A}^\top = \mathbf{A} \in \mathbb{R}^{n \times n}$, assume that its eigenvalues lie in $[0, 1]$ and are indexed in descending order, i.e.,

$$1 \geq \lambda_1 = \cdots = \lambda_p > \lambda_{p+1} \geq \cdots \geq \lambda_n \geq 0,$$

where p is the multiplicity of λ_1. Let the i-th eigengap of $\mathbf{A}$ be

$$\Delta_i \triangleq \lambda_i - \lambda_{i+1}.$$

Most of existing works handle only the case that $\Delta_1 = \lambda_1 - \lambda_2 > 0$, ignoring the cases that $\Delta_1 = 0$. All the cases are unified here via $\Delta_p > 0$ which holds always without loss of generality[2], i.e., $p < n$. Suppose that corresponding eigenvectors are $\mathbf{v}_1, \cdots, \mathbf{v}_n$, where $\mathbf{v}_1, \cdots, \mathbf{v}_p$ are the leading eigenvectors of $\mathbf{A}$. Let $\mathbf{V}_j = (\mathbf{v}_1, \cdots, \mathbf{v}_j)$. Our goal then is to find one of the leading eigenvectors, i.e.,

$$\mathbf{v} \in \mathcal{V}_{p,1} \triangleq \{\mathbf{v} \in \mathrm{span}(\mathbf{V}_p) : \|\mathbf{v}\|_2 = 1\}. \tag{8}$$

Denote $\mathbf{C}^{-1} = (\sigma\mathbf{I} - \mathbf{A})^{-1}$ as the shift-and-inverted matrix, where $\sigma > \lambda_1$. $\mathbf{C}^{-1}$'s eigenvalues then are $\mu_i = \frac{1}{\sigma - \lambda_i}$ satisfying

$$\mu_1 = \cdots = \mu_p > \mu_{p+1} \geq \cdots \geq \mu_n > 0,$$

while eigenvectors remain unchanged. Accordingly, define the i-th eigengap of $\mathbf{C}^{-1}$ as

$$\tau_i = \mu_i - \mu_{i+1}.$$

2. If $p = n$, the objective functions $\frac{1}{2}\mathbf{x}^\top \mathbf{A}\mathbf{x}$ and $\frac{1}{2}\mathbf{x}^\top \mathbf{C}^{-1}\mathbf{x}$ are almost constant and Problem (1)-(2) are trivial.

In particular, the relative eigengap of $\mathbf{C}^{-1}$ is

$$\frac{\tau_p}{\mu_1} = \frac{\mu_p - \mu_{p+1}}{\mu_1} = \frac{\frac{1}{\sigma-\lambda_p} - \frac{1}{\sigma-\lambda_{p+1}}}{\frac{1}{\sigma-\lambda_1}} = \frac{\Delta_p}{\sigma - \lambda_{p+1}}.$$

If $\mathbf{C}^{-1}$'s relative eigengap is larger than $\mathbf{A}$'s, i.e., $\frac{\tau_p}{\mu_1} > \frac{\Delta_p}{\lambda_1}$, then a faster rate can be achieved when shifting from Problem (1) to Problem (2), which is exactly the idea behind the shift-and-invert preconditioning. To this end, we follow Garber and Hazan (2015) and Wang et al. (2018)'s procedure (see Algorithm 7 in Appendix) to find a proper upper bound σ that is slightly larger than λ_1. Note that the procedure works regardless of how the eigengap is defined. Thus, Algorithm 7 is guaranteed to output

$$\sigma = \lambda_1 + c\Delta_p$$

where $c \in [\frac{1}{4}, \frac{3}{2}]$, as stated by the following theorem.

Theorem 1 *(Garber and Hazan, 2015; Wang et al., 2018) Let $\epsilon(\mathbf{x})$ be the function error with the least-squares subproblem. If the initial to final error ratio $\frac{\epsilon(\mathbf{x}_{init})}{\epsilon(\mathbf{x}_{final})}$ for the least-squares subproblems can be maintained as $\frac{32\cdot 10^{2m}+1}{\eta^{2m}}$ for the subproblem at Line 7, Algorithm 7 and $\frac{1024}{\eta^2}$ for the subproblem at Line 11, Algorithm 7, where $m = \lceil 8 \log \frac{16}{\|\mathbf{V}_p^\top \tilde{\mathbf{a}}_0\|_2^2} \rceil$, then we have the output $\sigma = \lambda_1 + c\Delta_p$ for certain $c \in [\frac{1}{4}, \frac{3}{2}]$ after $O(\log \frac{1}{\eta})$ iterations in the outer repeat-until loop.*

After Algorithm 7, we have σ satisfying

$$\frac{\tau_p}{\mu_1} = \frac{1}{c+1} \geq \frac{2}{5}, \tag{9}$$

and then can run the Riemannian gradient ascent to solve Problem (2). Let $h(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top \mathbf{C}^{-1}\mathbf{x}$. From Section 3, the Riemannian gradient ascent update can be written as

$$\begin{aligned} \mathbf{x}_{t+1} &= R(\mathbf{x}_t, \alpha_t \tilde{\nabla} h(\mathbf{x}_t)) \\ &= R(\mathbf{x}_t, \alpha_t (\mathbf{I} - \mathbf{x}_t\mathbf{x}_t^\top)\mathbf{C}^{-1}\mathbf{x}_t). \end{aligned} \tag{10}$$

As $(\frac{\mu_1}{\tau_p})^2 = O(1)$, the Riemannian gradient ascent takes only a logarithmic number of iterations $O(\log \frac{1}{\epsilon})$ to converge now, which does not have the quadratic dependence on $\frac{\lambda_1}{\lambda_1-\lambda_2}$ any more (Shamir, 2015, 2016a; Xu et al., 2017). In each iteration, however, we need to calculate the inverse-matrix-vector multiplication $\mathbf{C}^{-1}\mathbf{x}_t$. Instead of consuming a high cost to start from inverting $\mathbf{C}$, the multiplication can be directly approximated by solving the equivalent least-squares subproblem

$$\min_{\mathbf{x}} l_t(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top \mathbf{C}\mathbf{x} - \mathbf{x}_t^\top \mathbf{x} \tag{11}$$

to certain sub-optimality. The algorithmic steps are described in Algorithm 1, where only a few iterations are required for the warm-started least-squares solver to get the output $\widehat{\mathbf{C}^{-1}\mathbf{x}_t} \approx \mathbf{C}^{-1}\mathbf{x}_t = \arg\min_{\mathbf{x}} l_t(\mathbf{x})$. It is easy to see that the Riemannian gradient method recovers the shift-and-inverted power method, i.e., $\mathbf{x}_{t+1} = \frac{\mathbf{C}^{-1}\mathbf{x}_t}{\|\mathbf{C}^{-1}\mathbf{x}_t\|_2}$ using $\alpha_t = \frac{1}{\mathbf{x}_t^\top \mathbf{C}^{-1}\mathbf{x}_t}$ and $\mathbf{x}_{t+1} = \frac{\widehat{\mathbf{C}^{-1}\mathbf{x}_t}}{\|\widehat{\mathbf{C}^{-1}\mathbf{x}_t}\|_2}$ using $\alpha_t = \frac{1}{\mathbf{x}_t^\top \widehat{\mathbf{C}^{-1}\mathbf{x}_t}}$, in exact and inexact circumstances, respectively.

Algorithm 1 Shift-and-Invert Preconditioned Riemannian Gradient Eigensolver (SI-rgEIGS)

1: **Input:** matrix $\mathbf{A}$, shift σ, and initial $\mathbf{x}_0$, least-squares solver $\mathrm{ls}(\tilde{\mathbf{A}}, \tilde{\mathbf{b}}, \tilde{\mathbf{x}}_0)$ for solving $\min_{\tilde{\mathbf{x}}} \frac{1}{2}\tilde{\mathbf{x}}^\top \tilde{\mathbf{A}}\tilde{\mathbf{x}} - \tilde{\mathbf{b}}^\top \tilde{\mathbf{x}}$ with initial $\tilde{\mathbf{x}}_0$.
2: **for** $t = 0, 1, 2, \cdots$ **do**
3: $\widehat{\mathbf{C}^{-1}\mathbf{x}_t} \approx \mathrm{ls}(\mathbf{C}, \mathbf{x}_t, \frac{\mathbf{x}_t}{\mathbf{x}_t^\top \mathbf{C}\mathbf{x}_t})$
4: $\widehat{\tilde{\nabla} h(\mathbf{x}_t)} = (\mathbf{I} - \mathbf{x}_t\mathbf{x}_t^\top)\widehat{\mathbf{C}^{-1}\mathbf{x}_t}$
5: choose a step size $\alpha_t > 0$
6: set $\mathbf{x}_{t+1} = \frac{\mathbf{x}_t + \alpha_t \widehat{\tilde{\nabla} h(\mathbf{x}_t)}}{\|\mathbf{x}_t + \alpha_t \widehat{\tilde{\nabla} h(\mathbf{x}_t)}\|_2}$
7: terminate if stopping criterion is met
8: **end for**
9: **Output:** $\mathbf{x}_t$

4.2 Analysis

We would like to analyze the convergence of Riemannian gradient ascent for Problem (2). However, it turns out that directly analyzing Algorithm 1 would lead to a sub-optimal[3] convergence rate in terms of gap dependence. To overcome this issue, we make a change of variable, $\mathbf{y} = \mathbf{C}^{-1/2}\mathbf{x}$, in Problem (2), and then arrive at an equivalent form, i.e., Problem (3). Let $h(\mathbf{y}) = \frac{1}{2}\|\mathbf{y}\|_2^2$. From Section 3, the Riemannian gradient ascent update for Problem (3) can be written as

$$\begin{aligned} \mathbf{y}_{t+1} &= R(\mathbf{y}_t, \alpha_t \tilde{\nabla} h(\mathbf{y}_t)) \\ &= R(\mathbf{y}_t, \alpha_t (\mathbf{I} - \mathbf{y}_t\mathbf{y}_t^\top \mathbf{C})\mathbf{C}^{-1}\mathbf{y}_t), \end{aligned}$$

corresponding to Algorithm 2.

Algorithm 2 SI-rgEIGS for analysis

1: **Input:** matrix $\mathbf{A}$, shift σ, and initial $\mathbf{y}_0$, least-squares solver $\mathrm{ls}(\tilde{\mathbf{A}}, \tilde{\mathbf{b}}, \tilde{\mathbf{x}}_0)$ for solving $\min_{\tilde{\mathbf{x}}} \frac{1}{2}\tilde{\mathbf{x}}^\top \tilde{\mathbf{A}}\tilde{\mathbf{x}} - \tilde{\mathbf{b}}^\top \tilde{\mathbf{x}}$ with initial $\tilde{\mathbf{x}}_0$.
2: **for** $t = 0, 1, 2, \cdots$ **do**
3: $\widehat{\mathbf{C}^{-1}\mathbf{y}_t} \approx \mathrm{ls}(\mathbf{C}, \mathbf{y}_t, \|\mathbf{y}_t\|_2^2 \mathbf{y}_t)$
4: $\widehat{\tilde{\nabla} h(\mathbf{y}_t)} = (\mathbf{I} - \mathbf{y}_t\mathbf{y}_t^\top \mathbf{C})\widehat{\mathbf{C}^{-1}\mathbf{y}_t}$
5: choose a step size $\alpha_t > 0$
6: set $\mathbf{y}_{t+1} = \frac{\mathbf{y}_t + \alpha_t \widehat{\tilde{\nabla} h(\mathbf{y}_t)}}{\|\mathbf{y}_t + \alpha_t \widehat{\tilde{\nabla} h(\mathbf{y}_t)}\|_{\mathbf{C}}}$
7: terminate if stopping criterion is met
8: **end for**
9: **Output:** $\frac{\mathbf{y}_t}{\|\mathbf{y}_t\|_2}$

3. There will be an additional factor $\log \frac{\lambda_1}{\Delta_p}$ with the rate.

To uncover the true rate of the method, we analyze Algorithm 2 instead. For any $\mathbf{v} \in \mathcal{V}_{p,1}$ in (8), $\mathbf{C}^{-1/2}\mathbf{v}$ is an optimal solution to Problem (3). Noting that $\mathbf{C}^{-1/2}\mathbf{v} = \sqrt{\mu}_1\mathbf{v}$, it suffices to normalize the final $\mathbf{y}_T$ to get the solution instead of computing $\mathbf{C}^{1/2}\mathbf{y}_T$. Despite the convenience to analysis, Algorithm 2 is often not as efficient as Algorithm 1, due to the two more multiplications with $\mathbf{C}$ in Lines 4 and 6. We will experiment with both Algorithms for empirical studies in Section 6.

4.2.1 Potential Functions

To measure the progress of $\mathbf{C}^{1/2}\mathbf{y}_t$ to $\mathbf{V}_p$, we use a novel potential function defined by

$$\varphi(\mathbf{C}^{1/2}\mathbf{y}_t, \mathbf{V}_p) = -2\log\|\mathbf{V}_p^\top \mathbf{C}^{1/2}\mathbf{y}_t\|_2$$

for analysis. As

$$\|\mathbf{V}_p^\top \mathbf{C}^{1/2}\mathbf{y}_t\|_2 \leq \|\mathbf{V}_p\|_2\|\mathbf{y}_t\|_\mathbf{C} = 1,$$

we have

$$\varphi(\mathbf{C}^{1/2}\mathbf{y}_t, \mathbf{V}_p) \geq 0.$$

In fact,

$$\|\mathbf{V}_p^\top \mathbf{C}^{1/2}\mathbf{y}_t\|_2 = \cos\theta(\mathbf{C}^{1/2}\mathbf{y}_t, \mathbf{V}_p),$$

where $\theta(\mathbf{C}^{1/2}\mathbf{y}_t, \mathbf{V}_p) \in [0, \frac{\pi}{2}]$ represents the principal angle (Golub and Van Loan, 2013) between $\mathbf{C}^{1/2}\mathbf{y}_t$ and the space of the leading eigenvectors $\text{span}(\mathbf{V}_p)$. Particularly, it is worth noting that

$$\theta(\mathbf{C}^{1/2}\mathbf{y}_t, \mathbf{V}_p) = \min_{\mathbf{v}\in\text{span}(\mathbf{V}_p)} \theta(\mathbf{C}^{1/2}\mathbf{y}_t, \mathbf{v}),$$

where $\theta(\mathbf{C}^{1/2}\mathbf{y}_t, \mathbf{v}) \in [0, \frac{\pi}{2}]$. That is, the angle between a vector $\mathbf{z}$ and a p-dimensional subspace $\text{span}(\mathbf{V}_p)$ is equal to the minimum angle between $\mathbf{z}$ and any $\mathbf{v} \in \text{span}(\mathbf{V}_p)$. Thus, we can write that

$$\varphi(\mathbf{C}^{1/2}\mathbf{y}_t, \mathbf{V}_p) = \min_{\mathbf{v}\in\mathcal{V}_{p,1}} \varphi(\mathbf{C}^{1/2}\mathbf{y}_t, \mathbf{v}),$$

where

$$\varphi(\mathbf{C}^{1/2}\mathbf{y}_t, \mathbf{v}) = -2\log|\mathbf{v}^\top \mathbf{C}^{1/2}\mathbf{y}_t| = -2\log\cos\theta(\mathbf{C}^{1/2}\mathbf{y}_t, \mathbf{v})$$

for any $\mathbf{v} \in \mathcal{V}_{p,1}$. This property will help address the degeneracy in analysis. It is easy to see that if $\varphi(\mathbf{C}^{1/2}\mathbf{y}_t, \mathbf{V}_p)$ goes to 0, $\mathbf{C}^{1/2}\mathbf{y}_t$ must converge to certain vector $\mathbf{v} \in \mathcal{V}_{p,1}$. We also use the common potential function (Shamir, 2015)

$$\sin^2\theta(\mathbf{C}^{1/2}\mathbf{y}_t, \mathbf{V}_p) = 1 - \|\mathbf{V}_p^\top \mathbf{C}^{1/2}\mathbf{y}_t\|_2^2$$

to assist in our analysis. Two potential functions have the following connection:

$$\varphi(\mathbf{C}^{1/2}\mathbf{y}_t, \mathbf{V}_p) = -\log(1 - \sin^2\theta(\mathbf{C}^{1/2}\mathbf{y}_t, \mathbf{V}_p)) \geq \sin^2\theta(\mathbf{C}^{1/2}\mathbf{y}_t, \mathbf{V}_p).$$

It is easier to handle the normalization in Step 6 of Algorithm 2 by $\varphi(\mathbf{C}^{1/2}\mathbf{y}_t, \mathbf{V}_p)$, see, e.g., Eq. (19), than by $\sin^2\theta(\mathbf{C}^{1/2}\mathbf{y}_t, \mathbf{V}_p)$ which would lead to complicated fractional expressions (Shamir, 2015).

4.2.2 Main Results

Our main results then can be stated as follows.

Theorem 2 *Given a shift parameter $\sigma = \lambda_1 + c\Delta_p$ for $c \in [\frac{1}{4}, \frac{3}{2}]$, Algorithm 2 with fixed step-sizes and using accelerated gradient descent as the least-squares solver is able to find one of the leading eigenvectors of $\mathbf{A}$, i.e., $\varphi(\mathbf{C}^{1/2}\mathbf{y}_T, \mathbf{V}_p) < \epsilon$, after $T = O(\log \frac{1}{\epsilon})$ gradient steps, and the overall complexity is $O(\sqrt{\frac{\lambda_1}{\Delta_p}} \log \frac{1}{\epsilon})$.*

Remark The shift parameter can be guaranteed by Theorem 1. In practice, however, it is not necessary for us to use Algorithm 7 to get such a shift parameter. Instead, we introduce a user-friendly procedure for this purpose in Section 4.3. More importantly, in practice, we can leverage step-size schemes, which we view as an advantage of search methods over projection methods. For example, in experiments, we will use the popular Barzilai-Borwein (BB) step-size scheme which exploits the second-order information in a very cheap manner. It is simple and automatic without the need of line-search like backtracking or hand-tuning.

To prove the theorem, we need the following auxiliary lemmas whose proofs are deferred to Appendix.

Lemma 3 *For any $\mathbf{x} \in \mathrm{St}_{\mathbf{I}}(n, 1)$, it holds for any q that*

$$\lambda_1 - \mathbf{x}^\top \mathbf{A}\mathbf{x} \geq (\lambda_1 - \lambda_{q+1}) \sin^2 \theta(\mathbf{x}, \mathbf{V}_q).$$

This lemma reveals the connection between two measures of errors: objective function error and Chordal distance of subspaces (i.e., one of our two potential functions). In particular, applying the lemma to $\mathbf{C}^{-1}$ for $q = p$, one gets for $\mathbf{x} \in \mathrm{St}_{\mathbf{I}}(n, 1)$ that

$$\mu_1 - \mathbf{x}^\top \mathbf{C}^{-1}\mathbf{x} \geq (\tau_1 - \tau_{p+1}) \sin^2 \theta(\mathbf{x}, \mathbf{V}_p) = \tau_p \sin^2 \theta(\mathbf{x}, \mathbf{V}_p),$$

which now relates to a positive eigengap. This lemma can be used to handle the vanishing gap issue, e.g., $\lambda_1 = \lambda_2$.

Lemma 4 *For any $\mathbf{y} \in \mathrm{St}_{\mathbf{C}}(n, 1)$, it holds that*

$$\|\tilde{\nabla} h(\mathbf{y})\|_{\mathbf{C}}^2 \leq 4\mu_1^2 \sin^2 \theta(\mathbf{C}^{1/2}\mathbf{y}, \mathbf{V}_p). \tag{12}$$

This lemma bounds the gradient norm using one of the potential functions.

Lemma 5 *Let*

$$\epsilon_t(\mathbf{y}) = l_t(\mathbf{y}) - l_t(\mathbf{C}^{-1}\mathbf{y}_t) \text{ and } \xi_t = \widehat{\mathbf{C}^{-1}\mathbf{y}_t} - \mathbf{C}^{-1}\mathbf{y}_t.$$

- *We have that*

$$\begin{aligned} 2\epsilon_t(\mathbf{y}) &= \|\mathbf{y} - \mathbf{C}^{-1}\mathbf{y}_t\|_{\mathbf{C}}^2, \quad 2\epsilon_t(\widehat{\mathbf{C}^{-1}\mathbf{y}_t}) = \|\xi_t\|_{\mathbf{C}}^2, \\ 2\epsilon_t(\|\mathbf{y}_t\|_2^2 \mathbf{y}_t) &\leq \mu_1^2 \sin^2 \theta(\mathbf{C}^{1/2}\mathbf{y}_t, \mathbf{V}_p). \end{aligned}$$

- *If Nesterov's accelerated gradient descent is adopted for solving Problem (11) with warm-starter $\|\mathbf{y}_t\|_2^2 \mathbf{y}_t$, it can take $O(\sqrt{\frac{\lambda_1}{\Delta_p}} \log \frac{\epsilon_t(\|\mathbf{y}_t\|_2^2 \mathbf{y}_t)}{\epsilon_t(\widehat{\mathbf{C}^{-1}\mathbf{y}_t})})$ complexity to reach sub-optimality $\epsilon_t(\widehat{\mathbf{C}^{-1}\mathbf{y}_t})$.*

Since the least-squares solver for Problem (11) is warm-started, the initial error $\epsilon_t(\|\mathbf{y}_t\|_2^2\mathbf{y}_t)$ is much smaller than those from random initials. Other least-squares solvers are applicable as well, such as (accelerated) SVRG (Johnson and Zhang, 2013; Lin et al., 2015) for matrices with covariance structure and coordinate descent for matrices without covariance structure (Wang et al., 2018).

Proof of Theorem 2

The roadmap of the proof is that we first show the iteration complexity is $T = O(\log \frac{1}{\epsilon})$ by preconditioning (such that $\frac{\tau_p}{\mu_1} = O(1)$, see Eq. (9)) and appropriate size of the error in Riemannian gradients, and then prove with Lemma 5 the complexity of the subproblem in each iteration can be uniformly bounded by $O(\sqrt{\frac{\lambda_1}{\Delta_p}})$. The total complexity then naturally is $O(\sqrt{\frac{\lambda_1}{\Delta_p}}\log \frac{1}{\epsilon})$.

In order to obtain the iteration complexity, we need to derive a recurrence relation in our potential function $\varphi(\mathbf{C}^{1/2}\mathbf{y}_t, \mathbf{V}_p)$, see Eq. (20). To this end, we can analyze the potential function with the Riemannian gradient ascent update, i.e., Eq. (13), under the approximation error, i.e., Eq. (21). For brevity, denote

$$\theta_t \triangleq \theta(\mathbf{C}^{1/2}\mathbf{y}_t, \mathbf{V}_p) \text{ and } \varphi_t \triangleq \varphi(\mathbf{C}^{1/2}\mathbf{y}_t, \mathbf{V}_p)$$

throughout the proof. First, for any $\mathbf{v} \in \mathcal{V}_{p,1}$, by the update in Step 6 of Algorithm 2, we have that

$$\begin{aligned}\varphi(\mathbf{C}^{1/2}\mathbf{y}_{t+1}, \mathbf{v}) &= -2\log|\mathbf{v}^\top\mathbf{C}^{1/2}\mathbf{y}_{t+1}| \\ &= -2\log|\mathbf{v}^\top\mathbf{C}^{1/2}(\mathbf{y}_t + \alpha_t\widehat{\tilde{\nabla}h(\mathbf{y}_t)})| + 2\log\|\mathbf{y}_t + \alpha_t\widehat{\tilde{\nabla}h(\mathbf{y}_t)}\|_{\mathbf{C}}, \quad (13)\end{aligned}$$

where $\widehat{\tilde{\nabla}h(\mathbf{y}_t)}$ is inexact Riemannian gradient, i.e.,

$$\widehat{\tilde{\nabla}h(\mathbf{y}_t)} = (\mathbf{I} - \mathbf{y}_t\mathbf{y}_t^\top\mathbf{C})\widehat{\mathbf{C}^{-1}\mathbf{y}_t}.$$

From Lemma 5, we can write

$$\widehat{\tilde{\nabla}h(\mathbf{y}_t)} = \tilde{\nabla}h(\mathbf{y}_t) + (\mathbf{I} - \mathbf{y}_t\mathbf{y}_t^\top\mathbf{C})\xi_t,$$

where ξ_t is the error in approximating the inverse-matrix-vector multiplication in Line 3 of Algorithm 2 by a least-squares solver. We then can expand the first term in the above equation as follows

$$\begin{aligned}&|\mathbf{v}^\top\mathbf{C}^{1/2}(\mathbf{y}_t + \alpha_t\widehat{\tilde{\nabla}h(\mathbf{y}_t)})|^2 \\ =\ &|\mathbf{v}^\top\mathbf{C}^{1/2}(\mathbf{y}_t + \alpha_t\tilde{\nabla}h(\mathbf{y}_t)) + \alpha_t\mathbf{v}^\top\mathbf{C}^{1/2}(\mathbf{I} - \mathbf{y}_t\mathbf{y}_t^\top\mathbf{C})\xi_t|^2 \\ \geq\ &(|\mathbf{v}^\top\mathbf{C}^{1/2}(\mathbf{y}_t + \alpha_t\tilde{\nabla}h(\mathbf{y}_t)| - \alpha_t|\mathbf{v}^\top\mathbf{C}^{1/2}(\mathbf{I} - \mathbf{y}_t\mathbf{y}_t^\top\mathbf{C})\xi_t|)^2 \\ \geq\ &|\mathbf{v}^\top\mathbf{C}^{1/2}(\mathbf{y}_t + \alpha_t\tilde{\nabla}h(\mathbf{y}_t))|^2 - 2\alpha_t|\mathbf{v}^\top\mathbf{C}^{1/2}(\mathbf{y}_t + \alpha_t\tilde{\nabla}h(\mathbf{y}_t))||\mathbf{v}^\top\mathbf{C}^{1/2}(\mathbf{I} - \mathbf{y}_t\mathbf{y}_t^\top\mathbf{C})\xi_t| \\ \geq\ &|\mathbf{v}^\top\mathbf{C}^{1/2}(\mathbf{y}_t + \alpha_t\tilde{\nabla}h(\mathbf{y}_t))|^2(1 - 2\alpha_t\frac{|\mathbf{v}^\top\mathbf{C}^{1/2}(\mathbf{I} - \mathbf{y}_t\mathbf{y}_t^\top\mathbf{C})\xi_t|}{|\mathbf{v}^\top\mathbf{C}^{1/2}(\mathbf{y}_t + \alpha_t\tilde{\nabla}h(\mathbf{y}_t))|}). \quad (14)\end{aligned}$$

To proceed, we analyze both $|\mathbf{v}^\top \mathbf{C}^{1/2}(\mathbf{y}_t + \alpha_t \tilde{\nabla} h(\mathbf{y}_t))|$ and $|\mathbf{v}^\top \mathbf{C}^{1/2}(\mathbf{I} - \mathbf{y}_t \mathbf{y}_t^\top \mathbf{C})\xi_t|$ individually. For the former,

$$|\mathbf{v}^\top \mathbf{C}^{1/2}(\mathbf{y}_t + \alpha_t \tilde{\nabla} h(\mathbf{y}_t))| = |\mathbf{v}^\top \mathbf{C}^{1/2}\mathbf{y}_t + \alpha_t \mathbf{v}^\top \mathbf{C}^{1/2} \tilde{\nabla} h(\mathbf{y}_t)|,$$

where

$$\begin{aligned}
\mathbf{v}^\top \mathbf{C}^{1/2} \tilde{\nabla} h(\mathbf{y}_t) &= \mathbf{v}^\top \mathbf{C}^{1/2}\mathbf{C}^{-1}\mathbf{y}_t - \mathbf{v}^\top \mathbf{C}^{1/2}\mathbf{y}_t\mathbf{y}_t^\top \mathbf{C}\mathbf{C}^{-1}\mathbf{y}_t \\
&= \mathbf{v}^\top \mathbf{C}^{-1}\mathbf{C}^{1/2}\mathbf{y}_t - \mathbf{v}^\top \mathbf{C}^{1/2}\mathbf{y}_t\mathbf{y}_t^\top \mathbf{y}_t \\
&= \mu_1 \mathbf{v}^\top \mathbf{C}^{1/2}\mathbf{y}_t - \|\mathbf{y}_t\|_2^2 \mathbf{v}^\top \mathbf{C}^{1/2}\mathbf{y}_t \\
&= (\mu_1 - \|\mathbf{y}_t\|_2^2)\mathbf{v}^\top \mathbf{C}^{1/2}\mathbf{y}_t.
\end{aligned}$$

Since $\mathbf{y}_t \in \mathrm{St}_{\mathbf{C}}(n,1)$, it holds that

$$\mathbf{x}_t \triangleq \mathbf{C}^{1/2}\mathbf{y}_t \in \mathrm{St}_{\mathbf{I}}(n,1). \tag{15}$$

By Lemma 3, we then have

$$\mu_1 - \|\mathbf{y}_t\|_2^2 = \mu_1 - \mathbf{x}_t \mathbf{C}^{-1}\mathbf{x}_t \geq \tau_p \sin^2\theta(\mathbf{x}_t, \mathbf{V}_p).$$

Thus, it holds that

$$\begin{aligned}
|\mathbf{v}^\top \mathbf{C}^{1/2}(\mathbf{y}_t + \alpha_t \tilde{\nabla} h(\mathbf{y}_t))| &= (1 + \alpha_t(\mu_1 - \|\mathbf{y}_t\|_2^2))|\mathbf{v}^\top \mathbf{C}^{1/2}\mathbf{y}_t| \\
&\geq (1 + \alpha_t \tau_p \sin^2\theta(\mathbf{x}_t, \mathbf{V}_p))\cos\theta(\mathbf{x}_t, \mathbf{v}).
\end{aligned} \tag{16}$$

For the latter, by the Cauchy-Schwartz inequality,

$$\begin{aligned}
|\mathbf{v}^\top \mathbf{C}^{1/2}(\mathbf{I} - \mathbf{y}_t\mathbf{y}_t^\top \mathbf{C})\xi_t| &= |\mathbf{v}^\top \mathbf{C}^{1/2}(\mathbf{I} - \mathbf{y}_t\mathbf{y}_t^\top \mathbf{C})\mathbf{C}^{-1/2}\mathbf{C}^{1/2}\xi_t| \\
&\leq \|\mathbf{v}^\top \mathbf{C}^{1/2}(\mathbf{I} - \mathbf{y}_t\mathbf{y}_t^\top \mathbf{C})\mathbf{C}^{-1/2}\|_2 \|\xi_t\|_{\mathbf{C}}.
\end{aligned}$$

Letting $\mathbf{x}_t^\perp$ represents the orthogonal complement of $\mathbf{x}_t$, it holds that

$$\begin{aligned}
& \|\mathbf{v}^\top \mathbf{C}^{1/2}(\mathbf{I} - \mathbf{y}_t\mathbf{y}_t^\top \mathbf{C})\mathbf{C}^{-1/2}\|_2 \\
= \; & \|\mathbf{v}^\top (\mathbf{I} - \mathbf{C}^{1/2}\mathbf{y}_t\mathbf{y}_t^\top \mathbf{C}^{1/2})\|_2 \\
= \; & \|\mathbf{v}^\top (\mathbf{I} - \mathbf{x}_t\mathbf{x}_t^\top)\|_2 = \|\mathbf{v}^\top \mathbf{x}_t^\perp (\mathbf{x}_t^\perp)^\top\|_2 \\
= \; & \|\mathbf{v}^\top \mathbf{x}_t^\perp\|_2 = (\mathbf{v}^\top \mathbf{x}_t^\perp (\mathbf{x}_t^\perp)^\top \mathbf{v})^{1/2} \\
= \; & (\mathbf{v}^\top (\mathbf{I} - \mathbf{x}_t\mathbf{x}_t^\top)\mathbf{v})^{1/2} = (1 - (\mathbf{v}^\top \mathbf{x}_t)^2)^{1/2} \\
= \; & \sin\theta(\mathbf{x}_t, \mathbf{v}),
\end{aligned}$$

where the fourth equality is due to the orthogonal invariance of the 2-norm. Thus, we have that

$$|\mathbf{v}^\top \mathbf{C}^{1/2}(\mathbf{I} - \mathbf{y}_t\mathbf{y}_t^\top \mathbf{C})\xi_t| \leq \|\xi_t\|_{\mathbf{C}} \sin\theta(\mathbf{x}_t, \mathbf{v}), \tag{17}$$

By Equations (16)-(17), the first term in Equation (13) can be bounded as follows:

$$\begin{aligned} -2\log|\mathbf{v}^\top\mathbf{C}^{1/2}\mathbf{y}_{t+1}| &\leq -2\log|\mathbf{v}^\top\mathbf{C}^{1/2}\mathbf{y}_t| - 2\log(1+\alpha_t\tau_p\sin^2\theta(\mathbf{x}_t,\mathbf{V}_p)) \\ &\quad -\log(1-\frac{2\alpha_t\|\xi_t\|_{\mathbf{C}}\tan\theta(\mathbf{x}_t,\mathbf{v})}{1+\alpha_t\tau_p\sin^2\theta(\mathbf{x}_t,\mathbf{V}_p)}). \end{aligned} \tag{18}$$

For the second term in Equation (13),

$$\begin{aligned} \|\mathbf{y}_t+\alpha_t\widehat{\tilde{\nabla}h(\mathbf{y}_t)}\|_{\mathbf{C}}^2 &= (\mathbf{y}_t+\alpha_t\widehat{\tilde{\nabla}h(\mathbf{y}_t)})^\top\mathbf{C}(\mathbf{y}_t+\alpha_t\widehat{\tilde{\nabla}h(\mathbf{y}_t)}) \\ &= 1+2\alpha_t\mathbf{y}_t^\top\mathbf{C}\widehat{\tilde{\nabla}h(\mathbf{y}_t)}+\alpha_t^2(\widehat{\tilde{\nabla}h(\mathbf{y}_t)})^\top\mathbf{C}(\widehat{\tilde{\nabla}h(\mathbf{y}_t)}) \\ &= 1+\alpha_t^2\|\widehat{\tilde{\nabla}h(\mathbf{y}_t)}\|_{\mathbf{C}}^2, \end{aligned}$$

where we have used that $\mathbf{y}_t\in\mathrm{St}_{\mathbf{C}}(n,1)$ and thus

$$\begin{aligned} \mathbf{y}_t^\top\mathbf{C}\widehat{\tilde{\nabla}h(\mathbf{y}_t)} &= \mathbf{y}_t^\top\mathbf{C}(\mathbf{I}-\mathbf{y}_t\mathbf{y}_t^\top\mathbf{C})\widehat{\mathbf{C}^{-1}\mathbf{y}_t} \\ &= \mathbf{0}. \end{aligned}$$

Moreover,

$$\begin{aligned} \|\widehat{\tilde{\nabla}h(\mathbf{y}_t)}\|_{\mathbf{C}}^2 &= \|\tilde{\nabla}h(\mathbf{y}_t)+(\mathbf{I}-\mathbf{y}_t\mathbf{y}_t^\top\mathbf{C})\xi_t\|_{\mathbf{C}}^2 \\ &\leq 2(\|\tilde{\nabla}h(\mathbf{y}_t)\|_{\mathbf{C}}^2+\|\mathbf{C}^{1/2}(\mathbf{I}-\mathbf{y}_t\mathbf{y}_t^\top\mathbf{C})\xi_t\|_2^2) \\ &= 2(\|\tilde{\nabla}h(\mathbf{y}_t)\|_{\mathbf{C}}^2+\|\mathbf{C}^{1/2}(\mathbf{I}-\mathbf{y}_t\mathbf{y}_t^\top\mathbf{C})\mathbf{C}^{-1/2}\mathbf{C}^{1/2}\xi_t\|_2^2) \\ &\leq 2(\|\tilde{\nabla}h(\mathbf{y}_t)\|_{\mathbf{C}}^2+\|\mathbf{C}^{1/2}(\mathbf{I}-\mathbf{y}_t\mathbf{y}_t^\top\mathbf{C})\mathbf{C}^{-1/2}\|_2^2\|\xi_t\|_{\mathbf{C}}^2) \\ &= 2(\|\tilde{\nabla}h(\mathbf{y}_t)\|_{\mathbf{C}}^2+\|\xi_t\|_{\mathbf{C}}^2) \\ &\leq 2(4\mu_1^2\sin^2\theta(\mathbf{x}_t,\mathbf{V}_p)+\|\xi_t\|_{\mathbf{C}}^2), \end{aligned}$$

where the last inequality is by Lemma 4. Thus, it holds that

$$\begin{aligned} 2\log\|\mathbf{y}_t+\alpha_t\widehat{\tilde{\nabla}h(\mathbf{y}_t)}\|_{\mathbf{C}} &\leq \log(1+2\alpha_t^2(4\mu_1^2\sin^2\theta(\mathbf{x}_t,\mathbf{V}_p)+\|\xi_t\|_{\mathbf{C}}^2)) \\ &\leq 2\alpha_t^2(4\mu_1^2\sin^2\theta(\mathbf{x}_t,\mathbf{V}_p)+\|\xi_t\|_{\mathbf{C}}^2). \end{aligned} \tag{19}$$

By Equations (13),(18)-(19), it holds that

$$\begin{aligned} \varphi(\mathbf{C}^{1/2}\mathbf{y}_{t+1},\mathbf{v}) &\leq \varphi(\mathbf{C}^{1/2}\mathbf{y}_t,\mathbf{v})-2\log(1+\alpha_t\tau_p\sin^2\theta(\mathbf{x}_t,\mathbf{V}_p)) \\ &\quad -\log(1-\frac{2\alpha_t\|\xi_t\|_{\mathbf{C}}\tan\theta(\mathbf{x}_t,\mathbf{v})}{1+\alpha_t\tau_p\sin^2\theta(\mathbf{x}_t,\mathbf{V}_p)}) \\ &\quad +2\alpha_t^2(4\mu_1^2\sin^2\theta(\mathbf{x}_t,\mathbf{V}_p)+\|\xi_t\|_{\mathbf{C}}^2), \end{aligned}$$

for any $\mathbf{v} \in \mathcal{V}_{p,1}$. Taking the minimum with respect to $\mathbf{v}$ on both sides of the above inequality and noting the properties of the potential functions in Section 4.2.1, we arrive at

$$\begin{aligned}\varphi(\mathbf{C}^{1/2}\mathbf{y}_{t+1}, \mathbf{V}_p) &\leq \varphi(\mathbf{C}^{1/2}\mathbf{y}_t, \mathbf{V}_p) - 2\log(1+\alpha_t\tau_p\sin^2\theta(\mathbf{x}_t, \mathbf{V}_p)) \\ &\quad - \log(1 - \frac{2\alpha_t\|\xi_t\|_{\mathbf{C}}\tan\theta(\mathbf{x}_t, \mathbf{V}_p)}{1+\alpha_t\tau_p\sin^2\theta(\mathbf{x}_t, \mathbf{V}_p)}) \\ &\quad + 2\alpha_t^2(4\mu_1^2\sin^2\theta(\mathbf{x}_t, \mathbf{V}_p) + \|\xi_t\|_{\mathbf{C}}^2).\end{aligned}$$

Noting that $\|\xi_t\|_{\mathbf{C}}^2 = 2\epsilon_t(\widehat{\mathbf{C}^{-1}\mathbf{y}_t})$ by Lemma 5, one gets that

$$\begin{aligned}\varphi_{t+1} &\leq \varphi_t - 2\log(1+\alpha_t\tau_p\sin^2\theta_t) - \log(1 - \frac{2\alpha_t\sqrt{2\epsilon_t(\widehat{\mathbf{C}^{-1}\mathbf{y}_t})}\tan\theta_t}{1+\alpha_t\tau_p\sin^2\theta_t}) \\ &\quad + 2\alpha_t^2(4\mu_1^2\sin^2\theta_t + 2\epsilon_t(\widehat{\mathbf{C}^{-1}\mathbf{y}_t})). \qquad (20)\end{aligned}$$

We now consider the constant step-size setting, i.e., $\alpha_t \equiv \alpha > 0$. Let

$$\epsilon_t(\widehat{\mathbf{C}^{-1}\mathbf{y}_t}) = \frac{\tau_p^2}{32}\sin^2(2\theta_t). \qquad (21)$$

Plugging this expression into Equation (20), we have that

$$\begin{aligned}\varphi_{t+1} &\leq \varphi_t - 2\log(1+\alpha\tau_p\sin^2\theta_t) - \log(1 - \frac{2\alpha\sqrt{\frac{1}{16}\tau_p^2\sin^2(2\theta_t)}\tan\theta_t}{1+\alpha\tau_p\sin^2\theta_t}) \\ &\quad + 2\alpha^2(4\mu_1^2\sin^2\theta_t + \frac{1}{16}\tau_p^2\sin^2(2\theta_t)) \\ &= \varphi_t - 2\log(1+\alpha\tau_p\sin^2\theta_t) - \log(1 - \frac{\alpha\tau_p\sin^2\theta_t}{1+\alpha\tau_p\sin^2\theta_t}) \\ &\quad + 2\alpha^2(4\mu_1^2\sin^2\theta_t + \frac{1}{4}\tau_p^2\sin\theta_t\cos^2\theta_t) \\ &\leq \varphi_t - \log(1+\alpha\tau_p\sin^2\theta_t) + 9\alpha^2\mu_1^2\sin^2\theta_t \\ &\leq \varphi_t - \frac{\alpha\tau_p\sin^2\theta_t}{1+\alpha\tau_p\sin^2\theta_t} + 9\alpha^2\mu_1^2\sin^2\theta_t \\ &\leq \varphi_t - \alpha(\frac{\tau_p}{1+\alpha\tau_p} - 9\alpha\mu_1^2)\sin^2\theta_t, \qquad (22)\end{aligned}$$

where the third inequality is due to inequality $\log(1+x) \geq \frac{x}{1+x}$. Assume that

$$\alpha\tau_p < \frac{1}{125(1+\alpha\tau_p)}.$$

By Equation (9), it holds that

$$\alpha\tau_p < \frac{1}{125(1+\alpha\tau_p)} \leq (\frac{\tau_p}{\mu_1})^2\frac{1}{20(1+\alpha\tau_p)}, \qquad (23)$$

and thus

$$9\alpha\mu_1^2 < \frac{\tau_p}{2(1+\alpha\tau_p)}.$$

Equation (22) thus implies that

$$\varphi_{t+1} \leq \varphi_t - \frac{\alpha\tau_p}{2(1+\alpha\tau_p)}\sin^2\theta_t < \varphi_t. \tag{24}$$

Noting the inequality $\frac{x}{-\log(1-x)} \geq \frac{1}{1-\log(1-x)}$, it holds that

$$\begin{aligned}\sin^2\theta_t &= \frac{\sin^2\theta_t}{-\log(1-\sin^2\theta_t)}\cdot\varphi_t \geq \frac{1}{1-\log(1-\sin^2\theta_t)}\cdot\varphi_t \\ &= \frac{\varphi_t}{1+\varphi_t} \geq \frac{\varphi_t}{1+\varphi_0}.\end{aligned}$$

Combining the two inequalities above, we get that

$$\varphi_{t+1} \leq (1 - \frac{\alpha\tau_p}{2(1+\alpha\tau_p)}\cdot\frac{1}{1+\varphi_0})\varphi_t,$$

and thus

$$\begin{aligned}\varphi_T &\leq (1 - \frac{\alpha\tau_p}{2(1+\alpha\tau_p)}\cdot\frac{1}{1+\varphi_0})^T\varphi_0 \\ &\leq \exp\{1 - T\cdot\frac{\alpha\tau_p}{2(1+\alpha\tau_p)}\cdot\frac{1}{1+\varphi_0}\}\varphi_0 \equiv \Xi.\end{aligned}$$

Solving $\Xi = \epsilon$ for T yields that

$$\begin{aligned}T &= \frac{2(1+\alpha\tau_p)(1+\varphi_0)}{\alpha\tau_p}\log\frac{\varphi_0}{\epsilon} \\ &= O(\frac{1}{\alpha\tau_p}\log\frac{\varphi_0}{\epsilon}) = O((\frac{\mu_1}{\tau_p})^2\log\frac{\varphi_0}{\epsilon}) = O(\log\frac{1}{\epsilon}),\end{aligned}$$

where the last two equalities are due to Equations (23) and (9), respectively.

On the other hand, in each iteration, by Lemma 5 and Equation (21) the complexity for computing $\widehat{\mathbf{C}^{-1}\mathbf{y}_t}$ is

$$\begin{aligned}&O(\sqrt{\frac{\lambda_1}{\Delta_p}}\log\frac{\epsilon_t(\|\mathbf{y}_t\|_2^2\mathbf{y}_t)}{\epsilon_t(\widehat{\mathbf{C}^{-1}\mathbf{y}_t})}) = O(\sqrt{\frac{\lambda_1}{\Delta_p}}\log\frac{\frac{\mu_1^2}{2}\sin^2\theta_t}{\frac{1}{32}\tau_p^2\sin^2(2\theta_t)}) \\ =\ &O(\sqrt{\frac{\lambda_1}{\Delta_p}}\log\frac{\mu_1^2\sin^2\theta_t}{\tau_p^2\sin^2\theta_t\cos^2\theta_t}) = O(\sqrt{\frac{\lambda_1}{\Delta_p}}(\log\frac{\mu_1}{\tau_p} + \varphi_t)) \\ =\ &O(\sqrt{\frac{\lambda_1}{\Delta_p}}(\log\frac{\mu_1}{\tau_p} + \varphi_0)) = O(\sqrt{\frac{\lambda_1}{\Delta_p}}),\end{aligned}$$

where the last two equalities are due to Equations (24) and (9), respectively. Therefore, the total complexity is $O(\sqrt{\frac{\lambda_1}{\Delta_p}}\log\frac{1}{\epsilon})$. ■

Before closing this section, we introduce one method from Zhou et al. (2006) to get an upper bound on the maximum eigenvalue of a matrix, i.e., the shift parameter σ.

4.3 Shift by Lanczos

It was noted in Zhou et al. (2006) that an effective upper bound on λ_1 can be obtained from a small number of steps of the standard Lanczos procedure (Saad, 2011), because Lanczos iterations often *approximate* extreme eigenvalues fast. An m-step Lanczos yields the following Lanczos decomposition:

$$\mathbf{AQ} = \mathbf{QT} + \mathbf{re}^\top,$$

where $\mathbf{Q} \in \mathrm{St}_{\mathbf{I}}(n, m)$ is a basis of the m-dimensional Krylov subspace, i.e., an m Lanczos basis, $\mathbf{T} \in \mathbb{R}^{k\times k}$ is a symmetric tridiagonal matrix, $\mathbf{r} \in \mathbb{R}^{n\times 1}$ represents the residual vector, and $\mathbf{e} \in \mathbb{R}^{m\times 1}$ represents a standard unit vector with $e_{11} = 1$ and $e_{i1} = 0$ for $i > 1$. Thus, we have that

$$\|\mathbf{AQ}\|_2 \leq \|\mathbf{QT}\|_2 + \|\mathbf{re}^\top\|_2 = \|\mathbf{T}\|_2 + \|\mathbf{r}\|_2,$$

where $\|\mathbf{T}\|_2 + \|\mathbf{r}\|_2$ will be an upper bound on $\lambda_1(\mathbf{A})$. The pseudo code of the procedure is given in Algorithm 3, i.e., Algorithm 4.3 in Zhou et al. (2006). Small values of m won't cause $\zeta = 0$ to occur and won't make the algorithm dominate the total time cost for solving the problem. However, the advantage for implementation is clear compared to Algorithm 7, as it is almost parameter-free.

Algorithm 3 Shift by Lanczos

Input: matrix $\mathbf{A}$, random initial $\mathbf{q}$, iteration number m
$\mathbf{q} \leftarrow \frac{\mathbf{q}}{\|\mathbf{q}\|_2}$, $\mathbf{r} = \mathbf{Aq}$
$\eta = \mathbf{r}^\top\mathbf{q}$, $\mathbf{r} \leftarrow \mathbf{r} - \eta\mathbf{q}$
$\mathbf{T}_{11} = \eta$
for $j = 2, \cdots, \min\{m, 20\}$ **do**
  $\zeta = \|\mathbf{r}\|_2$, $\mathbf{q}_0 = \mathbf{q}$
  $\mathbf{q} = \frac{1}{\zeta}\mathbf{r}$, $\mathbf{r} = \mathbf{Aq}$
  $\eta = \mathbf{r}^\top\mathbf{q}$, $\mathbf{r} \leftarrow \mathbf{r} - \eta\mathbf{q}$
  $\mathbf{r} \leftarrow \mathbf{r} - \zeta\mathbf{q}_0$
  $\mathbf{T}_{j,j-1} = \zeta$, $\mathbf{T}_{j-1,j} = \zeta$, $\mathbf{T}_{j,j} = \eta$
end for
Output: $\sigma = \|\mathbf{T}\|_2 + \|\mathbf{r}\|_2$

5. Generalized Eigenvector Computation

We now consider Problem (4) for the generalized eigenvector computation. The same setting as for the standard eigenvalues is adopted for the generalized eigenvalues, except that $(\lambda_i, \mathbf{v}_i)$, $i = 1, 2, \cdots, n$, are the generalized eigenpairs of $(\mathbf{A}, \mathbf{B})$, i.e.,

$$\mathbf{Av}_i = \lambda_i\mathbf{Bv}_i.$$

Our goal is to find one of the leading generalized eigenvectors, i.e.,

$$\mathbf{v} \in \mathcal{V}_{p,1} = \{\mathbf{v} \in \mathbf{V}_p : \|\mathbf{v}\|_{\mathbf{B}} = 1\}. \tag{25}$$

For ease of extension, we first convert Problem (4) to the form of Problem (1) by setting $\mathbf{x} \leftarrow \mathbf{B}^{1/2}\mathbf{x}$ and get that

$$\max_{\mathbf{x}\in\mathbb{R}^{n\times 1}:\|\mathbf{x}\|_2=1} \frac{1}{2}\mathbf{x}^\top \mathbf{B}^{-1/2}\mathbf{A}\mathbf{B}^{-1/2}\mathbf{x}.$$

Applying the preconditioning, we have the shift-and-inverted matrix

$$\mathbf{C}^{-1} = (\sigma\mathbf{I} - \mathbf{B}^{-1/2}\mathbf{A}\mathbf{B}^{-1/2})^{-1} = \mathbf{B}^{1/2}(\sigma\mathbf{B} - \mathbf{A})^{-1}\mathbf{B}^{1/2}$$

for $\sigma > \lambda_1$, using the same notations on its eigenvalues as before. Setting $\mathbf{x} \leftarrow \mathbf{B}^{-1/2}\mathbf{x}$ in Problem (2) with above $\mathbf{C}^{-1}$ leads to Problem (5).

To locate a proper shift parameter, there is a similar procedure (See Algorithm 8 in Appendix) (Wang et al., 2016), which has the following guarantee.

Theorem 6 *(Wang et al., 2016) If the final error for the least-squares subproblems can be maintained as $\tilde{\epsilon} \leq \frac{1}{3084}(\frac{\eta}{18})^{m-1}$ in Algorithm 8, where $m = \lceil 8\log\frac{16}{\|\mathbf{V}_p^\top \mathbf{B}\tilde{\mathbf{a}}_0\|_2^2}\rceil$, then we have the output $\sigma = \lambda_1 + c\Delta_p$ for certain $c \in [\frac{1}{4}, \frac{3}{2}]$ after $O(\log\frac{1}{\eta})$ iterations in the outer repeat-until loop.*

It then holds that

$$\frac{\tau_p}{\mu_1} = \frac{1}{c+1} \geq \frac{2}{5}.$$

Let $h(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top\mathbf{B}(\sigma\mathbf{B} - \mathbf{A})^{-1}\mathbf{B}\mathbf{x}$. The Riemannian gradient ascent update on $\mathrm{St}_{\mathbf{B}}(n, 1)$ will be

$$\begin{aligned}
\mathbf{x}_{t+1} &= R(\mathbf{x}_t, \alpha_t\tilde{\nabla}h(\mathbf{x}_t)) \\
&= R(\mathbf{x}_t, \alpha_t(\mathbf{B}^{-1} - \mathbf{x}_t\mathbf{x}_t^\top)\mathbf{B}(\sigma\mathbf{B} - \mathbf{A})^{-1}\mathbf{B}\mathbf{x}_t) \\
&= R(\mathbf{x}_t, \alpha_t(\mathbf{I} - \mathbf{x}_t\mathbf{x}_t^\top\mathbf{B})(\sigma\mathbf{B} - \mathbf{A})^{-1}\mathbf{B}\mathbf{x}_t),
\end{aligned}$$

where the inverse-matrix-vector multiplication $(\sigma\mathbf{B} - \mathbf{A})^{-1}\mathbf{B}\mathbf{x}_t$ can be approximated by solving the following least-squares subproblem:

$$\min_{\mathbf{x}} l_t(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top(\sigma\mathbf{B} - \mathbf{A})\mathbf{x} - \mathbf{x}_t^\top\mathbf{B}\mathbf{x} \tag{26}$$

to certain sub-optimality. The algorithmic steps are described in Algorithm 4, where $\widehat{\mathbf{g}}_t \approx (\sigma\mathbf{B} - \mathbf{A})^{-1}\mathbf{B}\mathbf{x}_t = \arg\min_{\mathbf{x}} l_t(\mathbf{x})$.

5.1 Analysis

In order to analyze the convergence of the Riemannian gradient ascent for Problem (5), we similarly turn to its equivalent form by making a change of variable $\mathbf{y} = \mathbf{B}^{-1/2}\mathbf{C}^{-1/2}\mathbf{B}^{1/2}\mathbf{x}$, i.e., Problem (6). Let $h(\mathbf{y}) = \frac{1}{2}\|\mathbf{y}\|_{\mathbf{B}}^2$. The Riemannian gradient ascent update can be written as

$$\begin{aligned}
\mathbf{y}_{t+1} &= R(\mathbf{y}_t, \alpha_t\tilde{\nabla}h(\mathbf{y}_t)) \\
&= R(\mathbf{y}_t, \alpha_t(\mathbf{I} - \mathbf{y}_t\mathbf{y}_t^\top(\sigma\mathbf{B} - \mathbf{A}))(\sigma\mathbf{B} - \mathbf{A})^{-1}\mathbf{B}\mathbf{y}_t),
\end{aligned}$$

Algorithm 4 Shift-and-Invert Preconditioned Riemannian Gradient Generalized-Eigensolver (SI-rgGenEIGS)

Input: matrix pair $(\mathbf{A}, \mathbf{B})$, shift σ, and initial $\mathbf{x}_0$, least-squares solver $\mathrm{ls}(\tilde{\mathbf{A}}, \tilde{\mathbf{b}}, \tilde{\mathbf{x}}_0)$ for solving $\min_{\tilde{\mathbf{x}}} \frac{1}{2}\tilde{\mathbf{x}}^\top \tilde{\mathbf{A}}\tilde{\mathbf{x}} - \tilde{\mathbf{b}}^\top \tilde{\mathbf{x}}$ with initial $\tilde{\mathbf{x}}_0$.
for $t = 0, 1, 2, \cdots$ **do**
  $\widehat{\mathbf{g}_t} \approx \mathrm{ls}(\sigma\mathbf{B} - \mathbf{A}, \mathbf{B}\mathbf{x}_t, \frac{\mathbf{x}_t^\top \mathbf{B}\mathbf{x}_t}{\mathbf{x}_t^\top(\sigma\mathbf{B}-\mathbf{A})\mathbf{x}_t}\mathbf{x}_t)$
  $\widehat{\tilde{\nabla} h(\mathbf{x}_t)} = (\mathbf{I} - \mathbf{x}_t\mathbf{x}_t^\top \mathbf{B})\widehat{\mathbf{g}_t}$
  choose a step size $\alpha_t > 0$
  set $\mathbf{x}_{t+1} = \frac{\mathbf{x}_t + \alpha_t \widehat{\tilde{\nabla} h(\mathbf{x}_t)}}{\|\mathbf{x}_t + \alpha_t \widehat{\tilde{\nabla} h(\mathbf{x}_t)}\|_{\mathbf{B}}}$
  terminate if stopping criterion is met
end for
Output: $\mathbf{x}_t$

Algorithm 5 SI-rgGenEIGS for analysis

Input: matrix pair $(\mathbf{A}, \mathbf{B})$, shift σ, and initial $\mathbf{y}_0$, least-squares solver $\mathrm{ls}(\tilde{\mathbf{A}}, \tilde{\mathbf{b}}, \tilde{\mathbf{x}}_0)$ for solving $\min_{\tilde{\mathbf{x}}} \frac{1}{2}\tilde{\mathbf{x}}^\top \tilde{\mathbf{A}}\tilde{\mathbf{x}} - \tilde{\mathbf{b}}^\top \tilde{\mathbf{x}}$ with initial $\tilde{\mathbf{x}}_0$.
for $t = 0, 1, 2, \cdots$ **do**
  $\widehat{\mathbf{g}_t} \approx \mathrm{ls}(\sigma\mathbf{B} - \mathbf{A}, \mathbf{B}\mathbf{y}_t, \|\mathbf{y}_t\|_{\mathbf{B}}^2 \mathbf{y}_t)$
  $\widehat{\tilde{\nabla} h(\mathbf{y}_t)} = (\mathbf{I} - \mathbf{y}_t\mathbf{y}_t^\top(\sigma\mathbf{B} - \mathbf{A}))\widehat{\mathbf{g}_t}$
  choose a step size $\alpha_t > 0$
  set $\mathbf{y}_{t+1} = \frac{\mathbf{y}_t + \alpha_t \widehat{\tilde{\nabla} h(\mathbf{y}_t)}}{\|\mathbf{y}_t + \alpha_t \widehat{\tilde{\nabla} h(\mathbf{y}_t)}\|_{\sigma\mathbf{B}-\mathbf{A}}}$
  terminate if stopping criterion is met
end for
Output: $\frac{\mathbf{y}_t}{\|\mathbf{y}_t\|_{\mathbf{B}}}$

corresponding to Algorithm 5. Note that for any $\mathbf{v} \in \mathcal{V}_{p,1}$ in (25),

$$\mathbf{B}^{-1/2}\mathbf{C}^{-1/2}\mathbf{B}^{1/2}\mathbf{v} = \mathbf{B}^{-1/2} \cdot \sqrt{\mu}_1 \mathbf{B}^{1/2}\mathbf{v} = \sqrt{\mu}_1 \mathbf{v}$$

is an optimal solution to Problem (6). Accordingly, the potential functions are

$$\begin{aligned} \varphi(\mathbf{B}^{-1/2}\mathbf{C}^{1/2}\mathbf{B}^{1/2}\mathbf{y}_t, \mathbf{V}_p) &= -2\log\|\mathbf{V}_p^\top \mathbf{B}^{1/2}\mathbf{C}^{1/2}\mathbf{B}^{1/2}\mathbf{y}_t\|_2 \\ &= \min_{\mathbf{v}\in\mathcal{V}_{p,1}} \varphi(\mathbf{B}^{-1/2}\mathbf{C}^{1/2}\mathbf{B}^{1/2}\mathbf{y}_t, \mathbf{v}) \end{aligned}$$

and

$$\sin^2\theta(\mathbf{B}^{-1/2}\mathbf{C}^{1/2}\mathbf{B}^{1/2}\mathbf{y}_t, \mathbf{V}_p) = 1 - \|\mathbf{V}_p^\top \mathbf{B}^{1/2}\mathbf{C}^{1/2}\mathbf{B}^{1/2}\mathbf{y}_t\|_2^2.$$

We have the following parallel theorem and lemmas. The proofs of the lemmas are deferred to Appendix.

Theorem 7 *Given the shift parameter $\sigma = \lambda_1 + c\Delta_p$ for $c \in [\frac{1}{4}, \frac{3}{2}]$, Algorithm 5 with fixed step-sizes and using accelerated gradient descent as the least-squares solver is able to find one of the leading generalized eigenvectors, i.e., $\varphi(\mathbf{B}^{-1/2}\mathbf{C}^{1/2}\mathbf{B}^{1/2}\mathbf{y}_t, \mathbf{V}_p) < \epsilon$, after $T = O(\log \frac{1}{\epsilon})$ gradient steps, and the overall complexity is $O(\sqrt{\kappa(\mathbf{B})\frac{\lambda_1}{\Delta_p}} \log \frac{1}{\epsilon})$.*

Lemma 8 *For any $\mathbf{x} \in \mathrm{St}_{\mathbf{B}}(n,1)$, it holds for any q that*

$$\lambda_1 - \mathbf{x}^\top \mathbf{A}\mathbf{x} \geq (\lambda_1 - \lambda_{q+1}) \sin^2 \theta(\mathbf{x}, \mathbf{V}_q).$$

Applying the lemma to $\mathbf{B}^{1/2}\mathbf{C}^{-1}\mathbf{B}^{1/2}$ for $q = p$, one gets for $\mathbf{x} \in \mathrm{St}_{\mathbf{B}}(n,1)$ that

$$\mu_1 - \mathbf{x}^\top \mathbf{B}^{1/2}\mathbf{C}^{-1}\mathbf{B}^{1/2}\mathbf{x} \geq (\tau_1 - \tau_{p+1}) \sin^2 \theta(\mathbf{x}, \mathbf{V}_p) = \tau_p \sin^2 \theta(\mathbf{x}, \mathbf{V}_p).$$

Lemma 9 *For any $\mathbf{x} \in \mathrm{St}_{\sigma\mathbf{B}-\mathbf{A}}(n,1)$, it holds that*

$$\|\tilde{\nabla} h(\mathbf{y})\|^2_{\sigma\mathbf{B}-\mathbf{A}} \leq 4\mu_1^2 \sin^2 \theta(\mathbf{B}^{-1/2}\mathbf{C}^{1/2}\mathbf{B}^{1/2}\mathbf{y}, \mathbf{V}_p). \tag{27}$$

Lemma 10 *Let*

$$\epsilon_t(\mathbf{y}) = l_t(\mathbf{y}) - l_t((\sigma\mathbf{B} - \mathbf{A})^{-1}\mathbf{B}\mathbf{y}_t) \text{ and } \xi_t = \widehat{\mathbf{g}}_t - (\sigma\mathbf{B} - \mathbf{A})^{-1}\mathbf{B}\mathbf{y}_t.$$

- *We have that*

$$\begin{aligned} 2\epsilon_t(\mathbf{y}) &= \|\mathbf{y} - (\sigma\mathbf{B} - \mathbf{A})^{-1}\mathbf{B}\mathbf{y}_t\|^2_{\sigma\mathbf{B}-\mathbf{A}}, \quad 2\epsilon_t(\widehat{\mathbf{g}}_t) = \|\xi_t\|^2_{\sigma\mathbf{B}-\mathbf{A}}, \\ 2\epsilon_t(\|\mathbf{y}_t\|^2_{\mathbf{B}}\mathbf{y}_t) &\leq \mu_1^2 \sin^2 \theta(\mathbf{B}^{-1/2}\mathbf{C}^{1/2}\mathbf{B}^{1/2}\mathbf{y}_t, \mathbf{V}_p). \end{aligned}$$

- *If Nesterov's accelerated gradient descent is adopted for solving Problem (11) with warm-starter $\|\mathbf{y}_t\|^2_{\mathbf{B}}\mathbf{y}_t$, it can take $O(\sqrt{\kappa_{\mathbf{B}}\frac{\lambda_1}{\Delta_p}} \log \frac{\epsilon_t(\|\mathbf{y}_t\|^2_{\mathbf{B}}\mathbf{y}_t)}{\epsilon_t(\widehat{\mathbf{g}}_t)})$ complexity to reach sub-optimality $\epsilon_t(\widehat{\mathbf{g}}_t)$.*

Proof of Theorem 7
We only give the sketch of the proof by following that of Theorem 2. For any $\mathbf{v} \in \mathcal{V}_{p,1}$, consider

$$\begin{aligned} &\varphi(\mathbf{B}^{-1/2}\mathbf{C}^{1/2}\mathbf{B}^{1/2}\mathbf{y}_{t+1}, \mathbf{v}) = -2\log|\mathbf{v}^\top \mathbf{B}^{1/2}\mathbf{C}^{1/2}\mathbf{B}^{1/2}\mathbf{y}_{t+1}| \\ = \;& -2\log|\mathbf{v}^\top \mathbf{B}^{1/2}\mathbf{C}^{1/2}\mathbf{B}^{1/2}(\mathbf{y}_t + \alpha_t \widehat{\tilde{\nabla} h(\mathbf{y}_t)})| + 2\log\|\mathbf{y}_t + \alpha_t \widehat{\tilde{\nabla} h(\mathbf{y}_t)}\|_{\sigma\mathbf{B}-\mathbf{A}}, \end{aligned}$$

where, by Lemma 10, we have

$$\begin{aligned} \widehat{\tilde{\nabla} h(\mathbf{y}_t)} &= (\mathbf{I} - \mathbf{y}_t\mathbf{y}_t^\top(\sigma\mathbf{B} - \mathbf{A}))\widehat{\mathbf{g}}_t \\ &= \tilde{\nabla} h(\mathbf{y}_t) + (\mathbf{I} - \mathbf{y}_t\mathbf{y}_t^\top(\sigma\mathbf{B} - \mathbf{A}))\xi_t. \end{aligned}$$

For the first term above, we can write that

$$\begin{aligned} &|\mathbf{v}^\top \mathbf{B}^{1/2}\mathbf{C}^{1/2}\mathbf{B}^{1/2}(\mathbf{y}_t + \alpha_t \widehat{\tilde{\nabla} h(\mathbf{y}_t)})|^2 \\ \geq \;& |\mathbf{v}^\top \mathbf{B}^{1/2}\mathbf{C}^{1/2}\mathbf{B}^{1/2}(\mathbf{y}_t + \alpha_t \tilde{\nabla} h(\mathbf{y}_t))|^2 \\ & (1 - 2\alpha_t \frac{|\mathbf{v}^\top \mathbf{B}^{1/2}\mathbf{C}^{1/2}\mathbf{B}^{1/2}(\mathbf{I} - \mathbf{y}_t\mathbf{y}_t^\top(\sigma\mathbf{B} - \mathbf{A}))\xi_t|}{|\mathbf{v}^\top \mathbf{B}^{1/2}\mathbf{C}^{1/2}\mathbf{B}^{1/2}(\mathbf{y}_t + \alpha_t \tilde{\nabla} h(\mathbf{y}_t))|}). \end{aligned} \tag{28}$$

Let

$$\mathbf{x}_t \triangleq \mathbf{B}^{-1/2}\mathbf{C}^{1/2}\mathbf{B}^{1/2}\mathbf{y}_t \in \mathrm{St}_{\mathbf{B}}(n,1). \tag{29}$$

One then obtains that

$$\begin{aligned}
& \mathbf{v}^\top\mathbf{B}^{1/2}\mathbf{C}^{1/2}\mathbf{B}^{1/2}\tilde{\nabla}h(\mathbf{y}_t) \\
= \;& \mathbf{v}^\top\mathbf{B}^{1/2}\mathbf{C}^{1/2}\mathbf{B}^{1/2}(\sigma\mathbf{B}-\mathbf{A})^{-1}\mathbf{B}\mathbf{y}_t - \mathbf{v}^\top\mathbf{B}^{1/2}\mathbf{C}^{1/2}\mathbf{B}^{1/2}\mathbf{y}_t\mathbf{y}_t^\top\mathbf{B}\mathbf{y}_t \\
= \;& \mathbf{v}^\top\mathbf{B}^{1/2}\mathbf{C}^{1/2}\mathbf{B}^{1/2}\mathbf{B}^{-1/2}\mathbf{C}^{-1}\mathbf{B}^{-1/2}\mathbf{B}\mathbf{y}_t - \|\mathbf{y}_t\|_{\mathbf{B}}^2\mathbf{v}^\top\mathbf{B}^{1/2}\mathbf{C}^{1/2}\mathbf{B}^{1/2}\mathbf{y}_t \\
= \;& \mathbf{v}^\top\mathbf{B}^{1/2}\mathbf{C}^{-1}\mathbf{C}^{1/2}\mathbf{B}^{1/2}\mathbf{y}_t - \|\mathbf{y}_t\|_{\mathbf{B}}^2\mathbf{v}^\top\mathbf{B}^{1/2}\mathbf{C}^{1/2}\mathbf{B}^{1/2}\mathbf{y}_t \\
= \;& \mu_1\mathbf{v}^\top\mathbf{B}^{1/2}\mathbf{C}^{1/2}\mathbf{B}^{1/2}\mathbf{y}_t - \|\mathbf{y}_t\|_{\mathbf{B}}^2\mathbf{v}^\top\mathbf{B}^{1/2}\mathbf{C}^{1/2}\mathbf{B}^{1/2}\mathbf{y}_t \\
= \;& (\mu_1 - \|\mathbf{y}_t\|_{\mathbf{B}}^2)\mathbf{v}^\top\mathbf{B}^{1/2}\mathbf{C}^{1/2}\mathbf{B}^{1/2}\mathbf{y}_t \\
= \;& (\mu_1 - \mathbf{x}_t^\top\mathbf{B}^{1/2}\mathbf{C}^{-1}\mathbf{B}^{1/2}\mathbf{x}_t)\mathbf{v}^\top\mathbf{B}\mathbf{x}_t \\
\geq \;& \tau_p \sin^2\theta(\mathbf{x}_t,\mathbf{V}_p)\cdot\mathbf{v}^\top\mathbf{B}\mathbf{x}_t,
\end{aligned}$$

where the inequality is by Lemma 8. In addition,

$$\begin{aligned}
& |\mathbf{v}^\top\mathbf{B}^{1/2}\mathbf{C}^{1/2}\mathbf{B}^{1/2}(\mathbf{I}-\mathbf{y}_t\mathbf{y}_t^\top(\sigma\mathbf{B}-\mathbf{A}))\xi_t| \\
= \;& |\mathbf{v}^\top\mathbf{B}^{1/2}\mathbf{C}^{1/2}\mathbf{B}^{1/2}(\mathbf{I}-\mathbf{y}_t\mathbf{y}_t^\top(\sigma\mathbf{B}-\mathbf{A}))(\sigma\mathbf{B}-\mathbf{A})^{-1/2}(\sigma\mathbf{B}-\mathbf{A})^{1/2}\xi_t| \\
\leq \;& \|\mathbf{v}^\top\mathbf{B}^{1/2}\mathbf{C}^{1/2}\mathbf{B}^{1/2}(\mathbf{I}-\mathbf{y}_t\mathbf{y}_t^\top(\sigma\mathbf{B}-\mathbf{A}))(\sigma\mathbf{B}-\mathbf{A})^{-1/2}\|_2\|\xi_t\|_{\sigma\mathbf{B}-\mathbf{A}} \\
= \;& \underbrace{\|\mathbf{v}^\top\mathbf{B}^{1/2}\mathbf{C}^{1/2}\mathbf{B}^{1/2}(\mathbf{I}-\mathbf{y}_t\mathbf{y}_t^\top(\sigma\mathbf{B}-\mathbf{A}))(\sigma\mathbf{B}-\mathbf{A})^{-1/2}\|_2}_{\triangleq\ \Omega}\|\xi_t\|_{\sigma\mathbf{B}-\mathbf{A}}.
\end{aligned}$$

Since $\mathbf{A}$ and $\mathbf{B}$ are not commutative, Ω can be simplified in the following manner:

$$\begin{aligned}
\Omega^2 &= \mathbf{v}^\top\mathbf{B}^{\frac{1}{2}}\mathbf{C}^{\frac{1}{2}}\mathbf{B}^{\frac{1}{2}}(\mathbf{I}-\mathbf{y}_t\mathbf{y}_t^\top(\sigma\mathbf{B}-\mathbf{A}))(\sigma\mathbf{B}-\mathbf{A})^{-1}(\mathbf{I}-\mathbf{y}_t\mathbf{y}_t^\top(\sigma\mathbf{B}-\mathbf{A}))^\top\mathbf{B}^{\frac{1}{2}}\mathbf{C}^{\frac{1}{2}}\mathbf{B}^{\frac{1}{2}}\mathbf{v} \\
&= \mathbf{v}^\top\mathbf{B}^{\frac{1}{2}}\mathbf{C}^{\frac{1}{2}}\mathbf{B}^{\frac{1}{2}}((\sigma\mathbf{B}-\mathbf{A})^{-1} - \mathbf{y}_t\mathbf{y}_t^\top - \mathbf{y}_t\mathbf{y}_t^\top + \mathbf{y}_t\mathbf{y}_t^\top(\sigma\mathbf{B}-\mathbf{A})\mathbf{y}_t\mathbf{y}_t^\top)\mathbf{B}^{\frac{1}{2}}\mathbf{C}^{\frac{1}{2}}\mathbf{B}^{\frac{1}{2}}\mathbf{v} \\
&= \mathbf{v}^\top\mathbf{B}^{\frac{1}{2}}\mathbf{C}^{\frac{1}{2}}\mathbf{B}^{\frac{1}{2}}((\sigma\mathbf{B}-\mathbf{A})^{-1} - \mathbf{y}_t\mathbf{y}_t^\top)\mathbf{B}^{\frac{1}{2}}\mathbf{C}^{\frac{1}{2}}\mathbf{B}^{\frac{1}{2}}\mathbf{v} \\
&= \mathbf{v}^\top\mathbf{B}^{\frac{1}{2}}(\mathbf{I} - \mathbf{C}^{\frac{1}{2}}\mathbf{B}^{\frac{1}{2}}\mathbf{y}_t\mathbf{y}_t^\top\mathbf{B}^{\frac{1}{2}}\mathbf{C}^{\frac{1}{2}})\mathbf{B}^{\frac{1}{2}}\mathbf{v} \\
&= 1 - (\mathbf{v}^\top\mathbf{B}\mathbf{x}_t)^2,
\end{aligned}$$

where the second equality is due to $\mathbf{y}_t \in \mathrm{St}_{\sigma\mathbf{B}-\mathbf{A}}(n,1)$, the fourth equality uses $\sigma\mathbf{B}-\mathbf{A} = \mathbf{B}^{\frac{1}{2}}\mathbf{C}\mathbf{B}^{\frac{1}{2}}$, and the last one is by Equation (29). Thus, we can bound the previous first term in Equation (28):

$$\begin{aligned}
& |\mathbf{v}^\top\mathbf{B}^{1/2}\mathbf{C}^{1/2}\mathbf{B}^{1/2}(\mathbf{y}_t + \alpha_t\widehat{\tilde{\nabla}h(\mathbf{y}_t)})|^2 \\
\geq \;& (1+\alpha_t\tau_p\sin^2\theta(\mathbf{x}_t,\mathbf{V}_p))^2(\mathbf{v}^\top\mathbf{B}\mathbf{x}_t)^2(1 - 2\alpha_t\frac{\|\xi_t\|_{\sigma\mathbf{B}-\mathbf{A}}(1-(\mathbf{v}^\top\mathbf{B}\mathbf{x}_t)^2)}{(1+\alpha_t\tau_p\sin^2\theta(\mathbf{x}_t,\mathbf{V}_p))|\mathbf{v}^\top\mathbf{B}\mathbf{x}_t|}).
\end{aligned}$$

For the second term of the potential function, it holds that

$$\begin{aligned}\|\mathbf{y}_t + \alpha_t \widehat{\tilde{\nabla} h(\mathbf{y}_t)}\|^2_{\sigma\mathbf{B}-\mathbf{A}} &= (\mathbf{y}_t + \alpha_t \widehat{\tilde{\nabla} h(\mathbf{y}_t)})^\top (\sigma\mathbf{B}-\mathbf{A})(\mathbf{y}_t + \alpha_t \widehat{\tilde{\nabla} h(\mathbf{y}_t)}) \\ &= 1 + 2\alpha_t \mathbf{y}_t^\top (\sigma\mathbf{B}-\mathbf{A}) \widehat{\tilde{\nabla} h(\mathbf{y}_t)} + \alpha_t^2 \|\widehat{\tilde{\nabla} h(\mathbf{y}_t)}\|^2_{\sigma\mathbf{B}-\mathbf{A}} \\ &= 1 + \alpha_t^2 \|\widehat{\tilde{\nabla} h(\mathbf{y}_t)}\|^2_{\sigma\mathbf{B}-\mathbf{A}},\end{aligned}$$

and

$$\begin{aligned}& \|\widehat{\tilde{\nabla} h(\mathbf{y}_t)}\|^2_{\sigma\mathbf{B}-\mathbf{A}} \\ =\ & \|\tilde{\nabla} h(\mathbf{y}_t) + (\mathbf{I} - \mathbf{y}_t\mathbf{y}_t^\top(\sigma\mathbf{B}-\mathbf{A}))\xi_t\|^2_{\sigma\mathbf{B}-\mathbf{A}} \\ \leq\ & 2\|\widehat{\tilde{\nabla} h(\mathbf{y}_t)}\|^2_{\sigma\mathbf{B}-\mathbf{A}} + 2\|(\sigma\mathbf{B}-\mathbf{A})^{1/2}(\mathbf{I} - \mathbf{y}_t\mathbf{y}_t^\top(\sigma\mathbf{B}-\mathbf{A}))\xi_t\|_2^2 \\ \leq\ & 2\|\widehat{\tilde{\nabla} h(\mathbf{y}_t)}\|^2_{\sigma\mathbf{B}-\mathbf{A}} + 2\|(\sigma\mathbf{B}-\mathbf{A})^{1/2}(\mathbf{I} - \mathbf{y}_t\mathbf{y}_t^\top(\sigma\mathbf{B}-\mathbf{A}))(\sigma\mathbf{B}-\mathbf{A})^{-1/2}\|_2\|\xi_t\|^2_{\sigma\mathbf{B}-\mathbf{A}} \\ =\ & 2\|\widehat{\tilde{\nabla} h(\mathbf{y}_t)}\|^2_{\sigma\mathbf{B}-\mathbf{A}} + 2\|\mathbf{I} - (\sigma\mathbf{B}-\mathbf{A})^{1/2}\mathbf{y}_t\mathbf{y}_t^\top(\sigma\mathbf{B}-\mathbf{A})^{1/2}\|_2\|\xi_t\|^2_{\sigma\mathbf{B}-\mathbf{A}} \\ \leq\ & 2\|\widehat{\tilde{\nabla} h(\mathbf{y}_t)}\|^2_{\sigma\mathbf{B}-\mathbf{A}} + 2\|\xi_t\|^2_{\sigma\mathbf{B}-\mathbf{A}}.\end{aligned}$$

Thus, the potential function can be bounded as follows:

$$\begin{aligned}\varphi_{t+1} \leq\ & \varphi_t - 2\log(1+\alpha_t\tau_p\sin^2\theta_t) \\ & + \log(1 - 2\alpha_t \frac{\|\xi_t\|_{\sigma\mathbf{B}-\mathbf{A}}\tan\theta_t}{1+\alpha_t\tau_p\sin^2\theta_t}) \\ & + 2\alpha_t^2(\|\widehat{\tilde{\nabla} h(\mathbf{y}_t)}\|^2_{\sigma\mathbf{B}-\mathbf{A}} + \|\xi_t\|^2_{\sigma\mathbf{B}-\mathbf{A}}).\end{aligned}$$

The remaining proof can proceed exactly as in the proof of Theorem 2 to get the results, using Lemmas 9-10. ■

5.2 CCA

Given two views of data, canonical correlation analysis is to find a low-dimensional data representation of each view such that the cross-view correlation is maximized in the dimension reduced space (Hotelling, 1936). Formally, let $\mathbf{X} \in \mathbb{R}^{d_x \times n}$ and $\mathbf{Y} \in \mathbb{R}^{d_y \times n}$ be two views of n objects[4]. Define auto-covariance and cross-covariance matrices as follows:

$$\begin{aligned}\mathbf{C}_{xx} &= \frac{1}{n}\mathbf{X}\mathbf{X}^\top + r_x\mathbf{I}, \\ \mathbf{C}_{yy} &= \frac{1}{n}\mathbf{Y}\mathbf{Y}^\top + r_y\mathbf{I}, \\ \mathbf{C}_{xy} &= \frac{1}{n}\mathbf{X}\mathbf{Y}^\top,\end{aligned}$$

4. Assume their rows are mean-centered.

where r_x, r_y are regularization parameters for avoiding ill-conditioned matrices. Canonical correlation $\rho = (\rho_1, \cdots, \rho_d)$ between $\mathbf{X}$ and $\mathbf{Y}$ and corresponding pair of canonical vectors

$$\{(\phi_i, \psi_i) \in \mathbb{R}^{d_x \times 1} \times \mathbb{R}^{d_y \times 1} : i = 1, \cdots, d\}$$

is recursively defined as $\rho_i = (\phi_i)^\top \mathbf{C}_{xy}(\psi_i)$, where $d = \min\{d_x, d_y\}$ and

$$(\phi_i, \psi_i) \in \arg \max_{\substack{\phi^\top \mathbf{C}_{xx}\phi=1, \phi^\top \mathbf{C}_{xx}\phi_j=0 \\ \psi^\top \mathbf{C}_{yy}\psi=1, \psi^\top \mathbf{C}_{yy}\psi_j=0 \\ j=1,\cdots,i-1}} \phi^\top \mathbf{C}_{xy}\psi,$$

for $i = 1, \cdots, d$. Clearly, $\rho_1 \geq \cdots \geq \rho_d$. In particular, the top-1 canonical subspace pair (ϕ_1, ψ_1) corresponding to ρ_1 constitutes one of solutions to the following succinct maximization program:

$$\rho_1 = \max_{\|\phi\|_{\mathbf{C}_{xx}}=\|\psi\|_{\mathbf{C}_{yy}}=1} \phi^\top \mathbf{C}_{xy}\psi.$$

By change of variable $\mathbf{x} = \frac{1}{\sqrt{2}} \begin{pmatrix} \phi \\ \psi \end{pmatrix}$ and setting

$$\mathbf{A} = \begin{pmatrix} & \mathbf{C}_{xy} \\ \mathbf{C}_{xy}^\top & \end{pmatrix}, \mathbf{B} = \begin{pmatrix} \mathbf{C}_{xx} & \\ & \mathbf{C}_{yy} \end{pmatrix},$$

it is equivalent (Zhang, 2015) to Problem (4). Thus, we can call Algorithm 4 to get that

$$\psi(\mathbf{x}_T, \frac{1}{\sqrt{2}} \begin{pmatrix} \phi_1, \cdots, \phi_p \\ \psi_1, \cdots, \psi_p \end{pmatrix}) < \frac{\epsilon}{2}.$$

Suppose $\mathbf{x}_T = (x_1, \cdots, x_{d_x+d_y})^\top$ and denote

$$\mathbf{x}_T^{(1)} = (x_1, \cdots, x_{d_x})^\top \text{ and } \mathbf{x}_T^{(2)} = (x_{d_x+1}, \cdots, x_{d_x+d_y})^\top.$$

Letting

$$\tilde{\mathbf{x}}_T^{(1)} = \frac{\mathbf{x}_T^{(1)}}{\|\mathbf{x}_T^{(1)}\|_{\mathbf{C}_{xx}}} \text{ and } \tilde{\mathbf{x}}_T^{(2)} = \frac{\mathbf{x}_T^{(2)}}{\|\mathbf{x}_T^{(2)}\|_{\mathbf{C}_{yy}}}, \tag{30}$$

we then have that

$$\begin{aligned} & \max\{\sin^2\theta(\tilde{\mathbf{x}}_T^{(1)}, (\phi_1, \cdots, \phi_p)), \sin^2\theta(\tilde{\mathbf{x}}_T^{(2)}, (\psi_1, \cdots, \psi_p))\} \\ = \; & \max\{1 - \|(\phi_1, \cdots, \phi_p)^\top \mathbf{C}_{xx}\tilde{\mathbf{x}}_T^{(1)}\|_2^2, 1 - \|(\psi_1, \cdots, \psi_p)^\top \mathbf{C}_{yy}\tilde{\mathbf{x}}_T^{(2)}\|_2^2\} \\ < \; & \epsilon, \end{aligned}$$

by the following lemma.

Lemma 11 *If* $\psi(\mathbf{x}_T, \frac{1}{\sqrt{2}} \begin{pmatrix} \phi_1, \cdots, \phi_p \\ \psi_1, \cdots, \psi_p \end{pmatrix}) < \frac{\epsilon}{2}$, *we then have that*

$$\max\{\sin^2\theta(\tilde{\mathbf{x}}_T^{(1)}, (\phi_1, \cdots, \phi_p)), \sin^2\theta(\tilde{\mathbf{x}}_T^{(2)}, (\psi_1, \cdots, \psi_p))\} < \epsilon.$$

Proof Let

$$\mathbf{\Phi}_p = (\phi_1, \cdots, \phi_p), \quad \mathbf{\Psi}_p = (\psi_1, \cdots, \psi_p)).$$

It is easy to see that

$$\sin^2\theta(\mathbf{x}_T, \frac{1}{\sqrt{2}}\begin{pmatrix}\mathbf{\Phi}_p \\ \mathbf{\Psi}_p\end{pmatrix}) < \psi(\mathbf{x}_T, \frac{1}{\sqrt{2}}\begin{pmatrix}\mathbf{\Phi}_p \\ \mathbf{\Psi}_p\end{pmatrix}) < \frac{\epsilon}{2}.$$

Noting that

$$\|\mathbf{B}^{\frac{1}{2}}\mathbf{x}\|_2^2 = \|\mathbf{C}_{xx}^{\frac{1}{2}}\mathbf{x}_T^{(1)}\|_2^2 + \|\mathbf{C}_{yy}^{\frac{1}{2}}\mathbf{x}_T^{(2)}\|_2^2 = 1$$

and by the Cauchy-Schwartz inequality, we have that

$$\begin{aligned}
& \|\mathbf{\Phi}_p^\top \mathbf{C}_{xx}\mathbf{x}_T^{(1)} + \mathbf{\Psi}_p^\top \mathbf{C}_{yy}\mathbf{x}_T^{(2)}\|_2 \\
\leq\ & \|\mathbf{\Phi}_p^\top \mathbf{C}_{xx}\mathbf{x}_T^{(1)}\|_2 + \|\mathbf{\Psi}_p^\top \mathbf{C}_{yy}\mathbf{x}_T^{(2)}\|_2 = \frac{\|\mathbf{\Phi}_p^\top \mathbf{C}_{xx}\mathbf{x}_T^{(1)}\|_2}{\|\mathbf{C}_{xx}^{\frac{1}{2}}\mathbf{x}_T^{(1)}\|_2}\|\mathbf{C}_{xx}^{\frac{1}{2}}\mathbf{x}_T^{(1)}\|_2 + \frac{\|\mathbf{\Psi}_p^\top \mathbf{C}_{yy}\mathbf{x}_T^{(2)}\|_2}{\|\mathbf{C}_{yy}^{\frac{1}{2}}\mathbf{x}_T^{(2)}\|_2}\|\mathbf{C}_{yy}^{\frac{1}{2}}\mathbf{x}_T^{(2)}\|_2 \\
\leq\ & \sqrt{(\frac{\|\mathbf{\Phi}_p^\top \mathbf{C}_{xx}\mathbf{x}_T^{(1)}\|_2}{\|\mathbf{C}_{xx}^{\frac{1}{2}}\mathbf{x}_T^{(1)}\|_2})^2 + (\frac{\|\mathbf{\Psi}_p^\top \mathbf{C}_{yy}\mathbf{x}_T^{(2)}\|_2}{\|\mathbf{C}_{yy}^{\frac{1}{2}}\mathbf{x}_T^{(2)}\|_2})^2} = \sqrt{\|\mathbf{\Phi}_p^\top \mathbf{C}_{xx}\tilde{\mathbf{x}}_T^{(1)}\|_2^2 + \|\mathbf{\Psi}_p^\top \mathbf{C}_{yy}\tilde{\mathbf{x}}_T^{(2)}\|_2^2}.
\end{aligned}$$

It hence holds that

$$\|\mathbf{\Phi}_p^\top \mathbf{C}_{xx}\tilde{\mathbf{x}}_T^{(1)}\|_2^2 + \|\mathbf{\Psi}_p^\top \mathbf{C}_{yy}\tilde{\mathbf{x}}_T^{(2)}\|_2^2 \geq \|\mathbf{\Phi}_p^\top \mathbf{C}_{xx}\mathbf{x}_T^{(1)} + \mathbf{\Psi}_p^\top \mathbf{C}_{yy}\mathbf{x}_T^{(2)}\|_2^2 \geq 2(1 - \frac{\epsilon}{2}) = 2 - \epsilon.$$

There must be $\|\mathbf{\Phi}_p^\top \mathbf{C}_{xx}\tilde{\mathbf{x}}_T^{(1)}\|_2^2 \geq 1 - \epsilon$, as $\|\mathbf{\Psi}_p^\top \mathbf{C}_{yy}\tilde{\mathbf{x}}_T^{(2)}\|_2^2 \leq 1$. Similarly, one gets that $\|\mathbf{\Psi}_p^\top \mathbf{C}_{yy}\tilde{\mathbf{x}}_T^{(2)}\|_2^2 \geq 1 - \epsilon$. ■

5.3 Shift by Lanczos for Matrix Pairs

We introduce here the Lanczos procedure for a matrix pair in order to produce an effective upper bound on λ_1 to be used as the shift parameter σ. The m-step Lanczos now can be written as:

$$\mathbf{AQ} = \mathbf{BQT} + \mathbf{Bre}^\top,$$

where $\mathbf{Q} \in \mathrm{St}_{\mathbf{B}}(n, m)$ is a $\mathbf{B}$-orthonormal basis of the m-dimensional Krylov subspace. Thus, we have that

$$\|\mathbf{B}^{-1/2}\mathbf{AB}^{-1/2} \cdot \mathbf{B}^{1/2}\mathbf{Q}\|_2 \leq \|\mathbf{B}^{1/2}\mathbf{QT}\|_2 + \|\mathbf{B}^{1/2}\mathbf{re}^\top\|_2 = \|\mathbf{T}\|_2 + \|\mathbf{r}\|_{\mathbf{B}},$$

where $\|\mathbf{T}\|_2 + \|\mathbf{r}\|_{\mathbf{B}}$ will be an upper bound on $\lambda_1(\mathbf{B}^{-1/2}\mathbf{AB}^{-1/2})$. The pseudo code of the procedure is given in Algorithm 6, i.e., Algorithm 9.1 in Saad (2011), where the inverse-matrix-vector multiplication is handled in the same way as before.

Algorithm 6 Shift by Lanczos with Matrix Pairs

Input: matrix pair $(\mathbf{A}, \mathbf{B})$, random initial $\mathbf{q}$, iteration number m, least-squares solver $\text{ls}(\tilde{\mathbf{A}}, \tilde{\mathbf{b}}, \tilde{\mathbf{x}}_0)$ for solving $\min_{\tilde{\mathbf{x}}} \frac{1}{2}\tilde{\mathbf{x}}^\top \tilde{\mathbf{A}}\tilde{\mathbf{x}} - \tilde{\mathbf{b}}^\top \tilde{\mathbf{x}}$ with initial $\tilde{\mathbf{x}}_0$.
$\mathbf{q} \leftarrow \frac{\mathbf{q}}{\|\mathbf{q}\|_{\mathbf{B}}}$, $\mathbf{r} = \mathbf{A}\mathbf{q}$
$\eta = \mathbf{r}^\top \mathbf{q}$, $\mathbf{r} \approx ls(\mathbf{B}, \mathbf{r}, \eta\mathbf{q})$, $\mathbf{r} \leftarrow \mathbf{r} - \eta\mathbf{q}$
$\mathbf{T}_{11} = \eta$
for $j = 2, \cdots, \min\{m, 20\}$ **do**
  $\zeta = \|\mathbf{r}\|_{\mathbf{B}}$, $\mathbf{q}_0 = \mathbf{q}$
  $\mathbf{q} = \frac{1}{\zeta}\mathbf{r}$, $\mathbf{r} = \mathbf{A}\mathbf{q}$
  $\eta = \mathbf{r}^\top \mathbf{q}$, $\mathbf{r} \approx ls(\mathbf{B}, \mathbf{r}, \eta\mathbf{q})$, $\mathbf{r} \leftarrow \mathbf{r} - \eta\mathbf{q}$
  $\mathbf{r} \leftarrow \mathbf{r} - \zeta\mathbf{q}_0$
  $\mathbf{T}_{j,j-1} = \zeta$, $\mathbf{T}_{j-1,j} = \zeta$, $\mathbf{T}_{j,j} = \eta$
end for
Output: $\sigma = \|\mathbf{T}\|_2 + \|\mathbf{r}\|_{\mathbf{B}}$

6. Experiments

We now test our proposed eigensolvers on both synthetic and real data. Throughout experiments, four iterations for the least-squares solvers are run to approximately solve those least-squares subproblems. Nesterov's accelerated gradient descent is used mostly for this purpose. Each experiment uses the same random initials $\mathbf{x}_0$ across eigensolvers unless otherwise stated. All the algorithms were implemented in matlab and running single threaded. All the ground-truth information was obtained by matlab's "eigs" function for evaluation purpose only.

6.1 Synthetic Data

In this section, we focus on synthetic data starting from standard eigenvalue problems. Following Shamir (2015), synthetic data is generated using $\mathbf{A}$'s full eigenvalue decomposition $\mathbf{A} = \mathbf{V}_n \mathbf{\Sigma} \mathbf{V}_n^\top$, where $\mathbf{\Sigma}$ is diagonal. Specifically, it suffices to generate random orthogonal matrix $\mathbf{V}_n$ and set $\mathbf{\Sigma} = \text{diag}(1, 1-\Delta_1, 1-1.1\Delta_1, \cdots, 1-1.4\Delta_1, g_1/n, \cdots, g_{n-6}/n)$ with g_i being standard normal samples, i.e., $g_i \sim \mathcal{N}(0,1)$. Here we set $n = 1000$ and $\sigma = 1.005$. Values of $\Delta_1 \in \{5 \times 10^{-3}, 5 \times 10^{-4}\}$ are used to generate two synthetic datasets. Three eigensolvers are compared: Rimennian gradient descent solver (rgEIGS), shift-and-invert preconditioned Rimennian gradient descent solver (SI-rgEIGS), and the shift-and-inverted power method (SI-PM) (Garber et al., 2016). Step-sizes are constant and hand-tuned for both rgEIGS and SI-rgEIGS. Two performance measures are used: relative function error $\frac{\lambda_1 - \mathbf{x}^\top \mathbf{A}\mathbf{x}}{\lambda_1}$ and potential $\sin^2\theta(\mathbf{x}, \mathbf{v}_1) = 1 - (\mathbf{x}^\top \mathbf{v}_1)^2$ for $\mathbf{x}$ satisfying $\|\mathbf{x}\|_2 = 1$. Smaller is better for both measures.

Figure 1 and Figure 2 show the convergence curves of three algorithms on the two synthetic datasets, in terms of the wall-clock time in seconds and counts of matrix vector multiplications (denoted as # $\mathbf{A}$*$\mathbf{x}$). We can see that two shift-and-invert preconditioned methods, i.e., SI-rgEIGS and SI-PM, outperforms the rgEIGS that is unpreconditioned. This demonstrates that the shift-and-invert preconditioning can accelerate Riemannian gradient

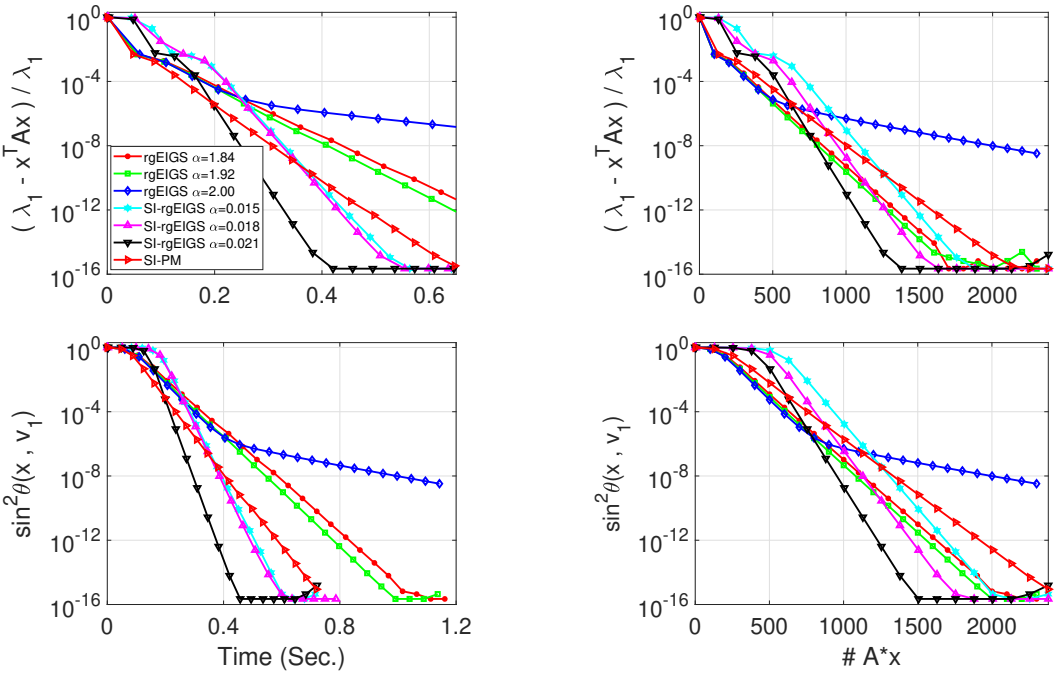

Figure 1: Algorithms for standard eigenvalue problems on synthetic data ($\Delta_1 = 5 \times 10^{-3}$).

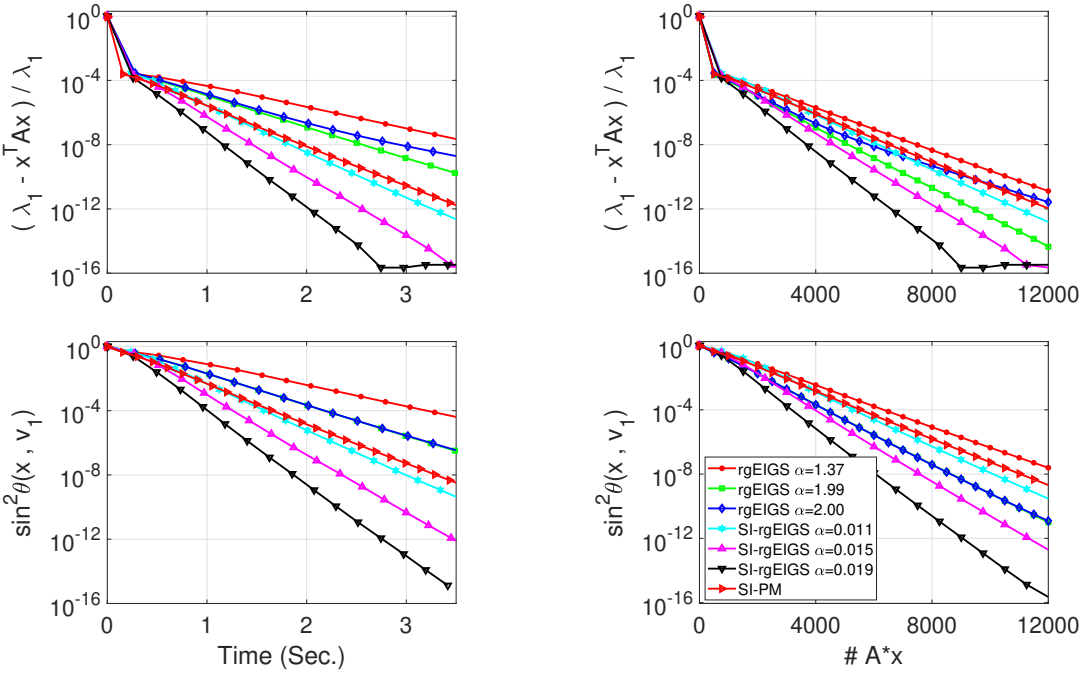

Figure 2: Algorithms for standard eigenvalue problems on synthetic data ($\Delta_1 = 5 \times 10^{-4}$).

methods for solving standard eigenvalue problems as well. In addition, SI-rgEIGS can run faster than SI-PM with proper step-sizes. This implies potential advantages of Riemannian eigensolvers in practice.

Before turning to generalized eigenvalue problems, we would like to see the influence of the shift parameter on the performance. Figure 3 reports the convergence of SI-rgEIGS under different shift parameters and best-tuned step-sizes. It shows that smaller values of σ can reap faster convergence, agreeing with the theory.

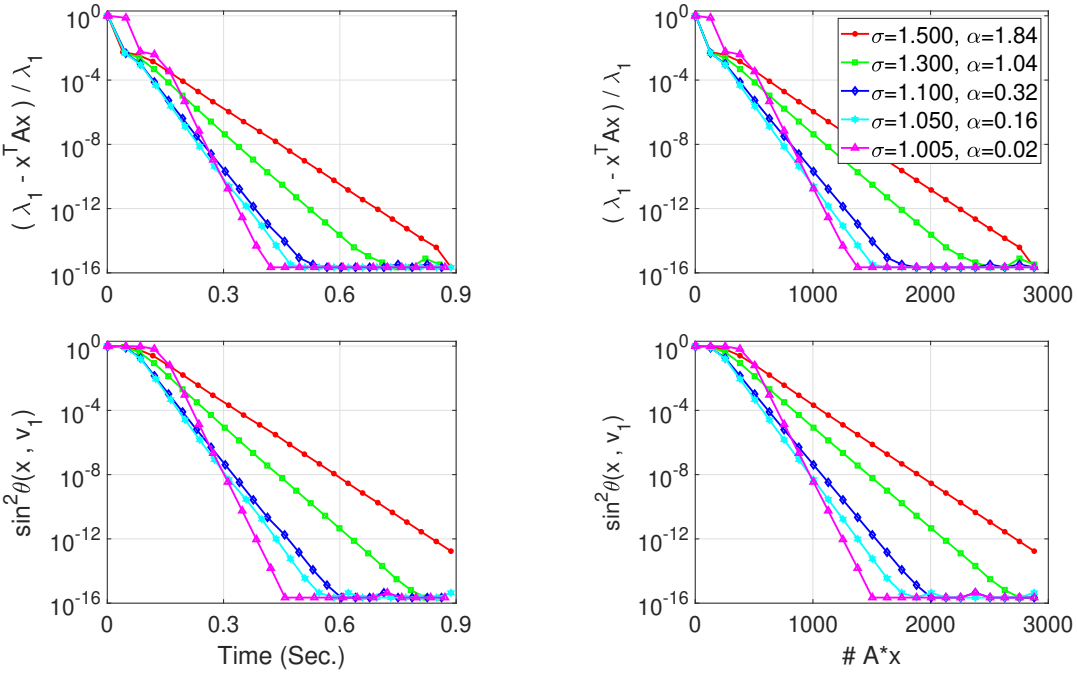

Figure 3: Influence of the shift parameter for SI-rgEIGS on synthetic data ($\Delta_1 = 5 \times 10^{-3}$).

For generalized eigenvalue problems, we generate one synthetic dataset using $\mathbf{A} = \mathbf{B}\mathbf{V}_n\mathbf{\Sigma}\mathbf{V}_n^\top\mathbf{B}$, where $\mathbf{\Sigma}$ is set the same as before with $\Delta_1 = 5 \times 10^{-3}$. First, random orthogonal matrix $\mathbf{U} \in \mathbb{R}^{n\times n}$ and diagonal matrix $\mathbf{D}$ are generated to get $\mathbf{B} = \mathbf{U}\mathbf{D}\mathbf{U}^\top$, where $\mathbf{D} = \mathrm{diag}(1 + g_1/n, \cdots, 1 + g_n/n)$ with g_i being standard normal samples. Then $\mathbf{B}$-orthogonal matrix $\mathbf{V}_n$, i.e., $\mathbf{V}_n\mathbf{B}\mathbf{V}_n^\top = \mathbf{I}$, is generated. Finally, $\mathbf{A}$ is obtained. Three generalized eigensolvers are compared accordingly: Rimennian gradient descent solver (rgGenEIGS), shift-and-invert preconditioned Rimennian gradient descent solver (SI-rgGenEIGS), and the shift-and-inverted power method (SI-PM-gen) (Wang et al., 2016). The potential now is $\sin^2\theta(\mathbf{x}, \mathbf{v}_1) = 1 - (\mathbf{x}^\top\mathbf{B}\mathbf{v}_1)^2$ for $\mathbf{x}$ satisfying $\|\mathbf{x}\|_\mathbf{B} = 1$. Following the same setting as for the standard case, the performance of three algorithms is reported in Figure 4. Similar patterns are observed.

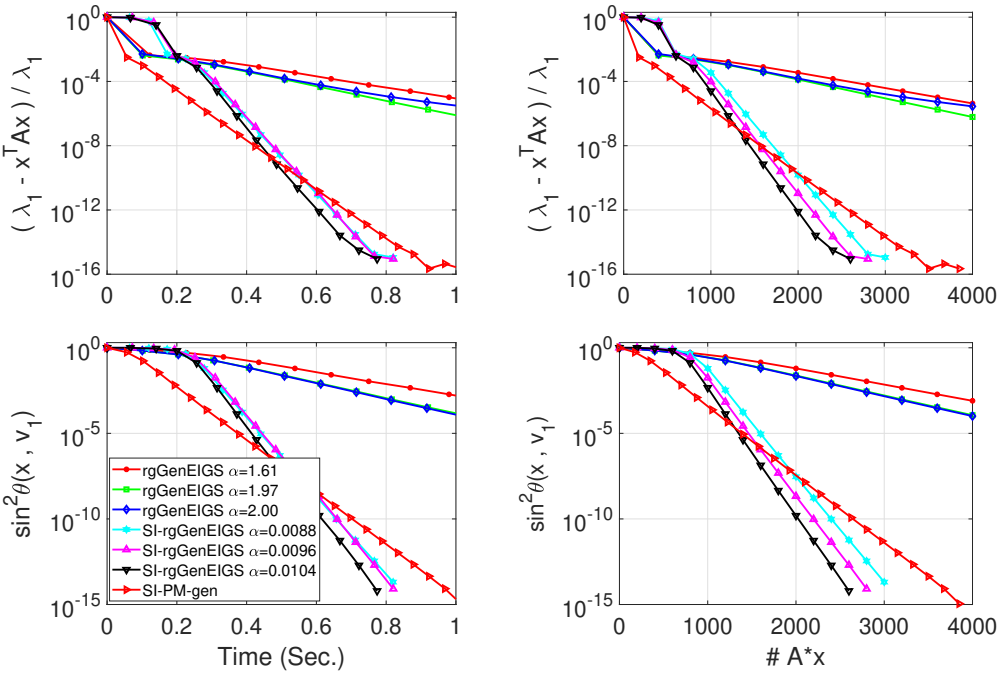

Figure 4: Algorithms for generalized eigenvalue problems on synthetic data ($\Delta_1 = 5\times10^{-3}$).

6.2 Real Data

We now demonstrate the performance of our algorithms on real data with $\Delta_1 = 5 \times 10^{-3}$.

Table 1: Statistics of the matrix data.

Matrix	n	# nonzero entries
hangGlider5	16011	155246
Boeing35	30237	1450163
indef_d	60000	299998
indef_a	60008	255004
dimacs10_ct	67578	336352
dimacs10_nv	84538	416998
ch7	17640	1816920

6.2.1 Standard Case

We download real data from the sparse matrix collection[5]. The statistics of the matrix data is given in Table 1. We compare our SI-rgEIGS (i.e., Algorithm 1) and SI-rgEIGS (F.A.)[6] (i.e., Algorithm 2) with the SI-PM (Garber et al., 2016), accelerated power method (Xu

5. `www.cise.ufl.edu/research/sparse/matrices/`
6. F.A. stands for "for analysis".

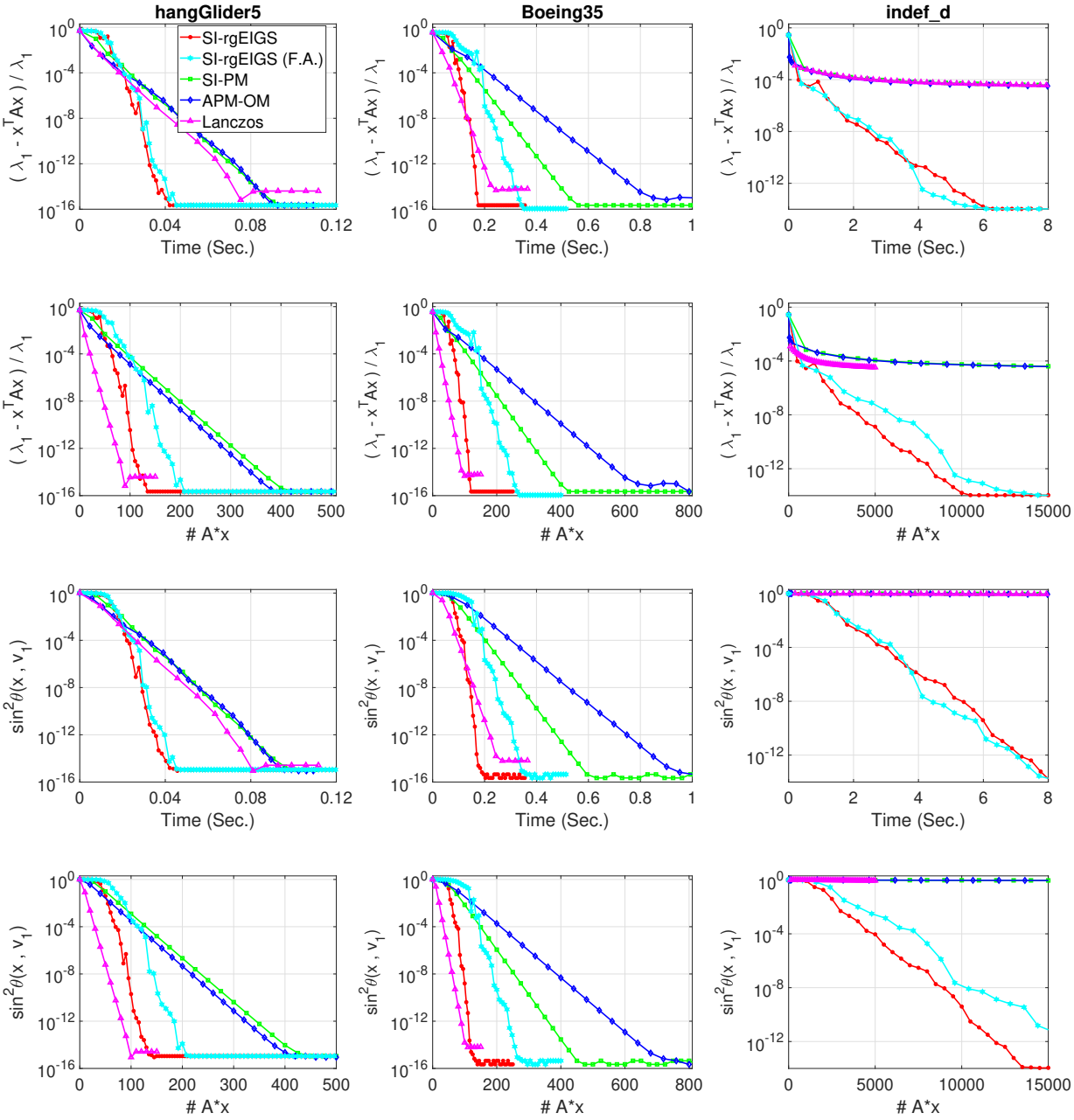

Figure 5: Algorithms for standard eigenvalue problems on real data - part I.

et al., 2018a) with optimal momentum $\beta = \lambda_2^2/4$, i.e., APM-OM, for short, as well as the symmetric Lanczos with thick restart (Watkins, 2007). All the three competitors are eigensolvers with acceleration. Particularly, the performance of the Lanczos algorithm relies heavily on a large amount of memory consumption which is significantly much larger than those needed by other considered algorithms here. In order for comparisons to be as fair as possible, we try to run the Lanczos algorithm with a minimum memory consumption[7].

7. This minimum memory consumption is still larger than those required by other algorithms.

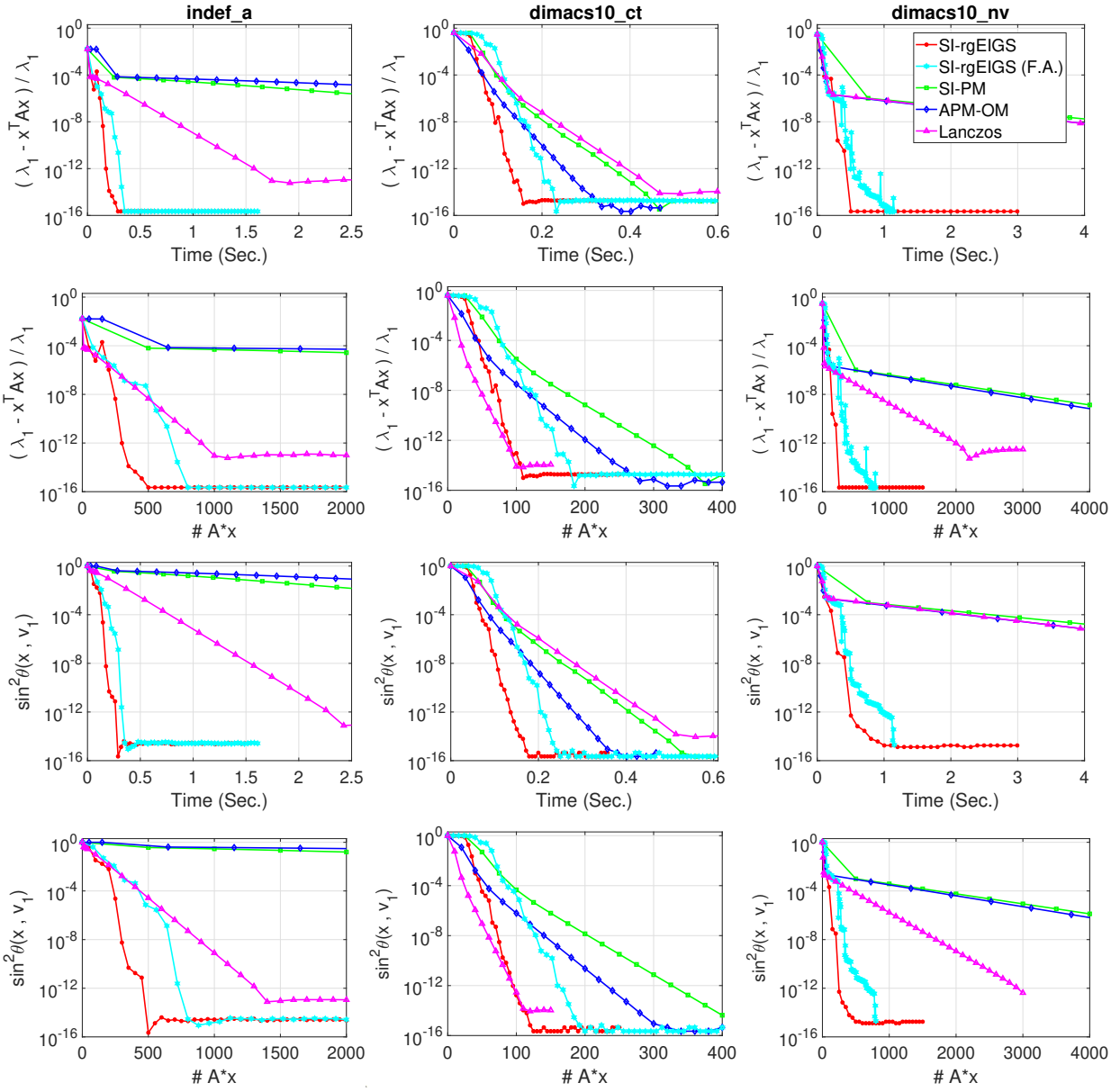

Figure 6: Algorithms for standard eigenvalue problems on real data - part II.

The shift parameter is obtained using Algorithm 3 with parameter $m = 9$ and used for all the shift-and-invert preconditioning based algorithms. To sidestep the difficult job of hand-tuning step-sizes for Algorithms 1-2, we use the following simplified Riemannian Barzilai-Borwein (BB) step-sizes (Barzilai and Borwein, 1988; Iannazzo and Porcelli, 2017),

$$\alpha_t = \frac{\|\mathbf{x}_t - \mathbf{x}_{t-1}\|_2^2}{|(\mathbf{x}_t - \mathbf{x}_{t-1})^\top (\widehat{\tilde{\nabla} h(\mathbf{x}_t)} - \widehat{\tilde{\nabla} h(\mathbf{x}_{t-1})})|} \quad \text{or} \quad \frac{|(\mathbf{x}_t - \mathbf{x}_{t-1})^\top (\widehat{\tilde{\nabla} h(\mathbf{x}_t)} - \widehat{\tilde{\nabla} h(\mathbf{x}_{t-1})})|}{\|\widehat{\tilde{\nabla} h(\mathbf{x}_t)} - \widehat{\tilde{\nabla} h(\mathbf{x}_{t-1})}\|_2^2}.$$

with initial step-sizes set to $\alpha_0 = 10^{-2}$. It is a non-monotone step-size scheme for which the values of quality measures are not necessarily monotonically decreasing. Note that inexact Riemannian gradients $\widehat{\tilde{\nabla} h(\mathbf{x}_t)}$ instead of exact ones $\tilde{\nabla} h(\mathbf{x})$ are used here. Nonetheless, our SI-rgEIGS with this step-size scheme still performs well and significantly better than the competitors in terms of the wall-clock time, and the SI-rgEIGS (F.A.) also outperform the competitors in most cases, as observed in Figure 5 and Figure 6. In a few cases (i.e., hangGlider5 and Boeing35), Lanczos is ranked the best in terms of the count of matrix vector multiplications. However, it is worth pointing out that a significant computational overhead is unreasonably omitted in the Lanczos process, such as reorthogonalization, Rayleigh-Rits procedure, and restart. That is why it is often not the case in terms of the running time.

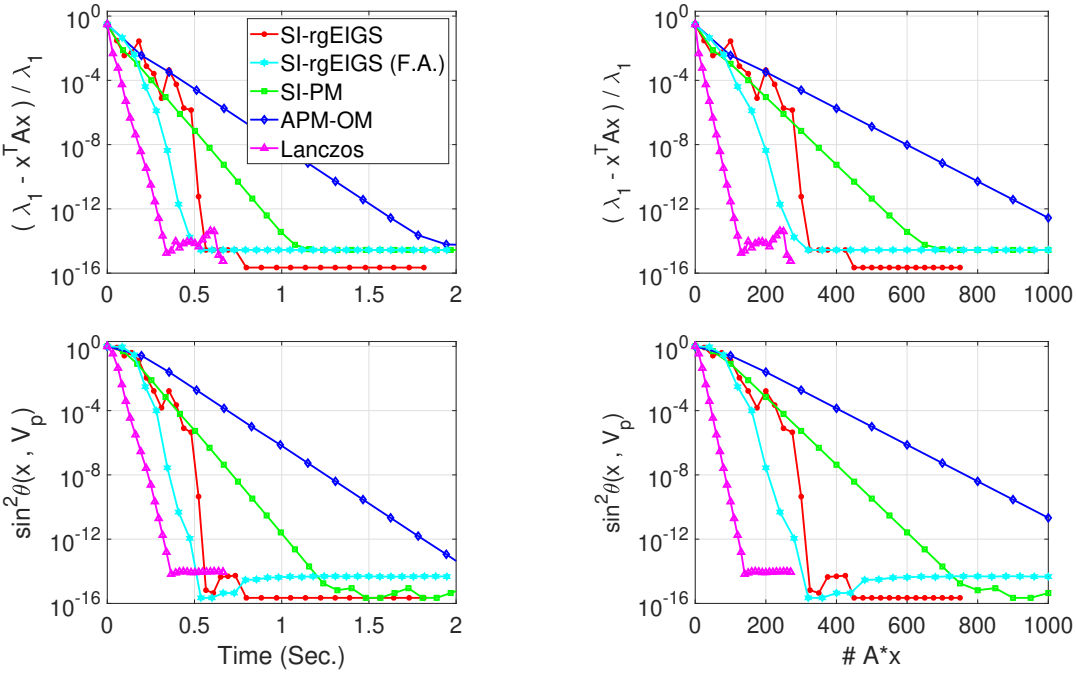

Figure 7: Algorithms for standard eigenvalue problems on real data ch7.

Figure 7 for ch7 is a special case. The underlying matrix is $\mathbf{A} = \mathbf{R}\mathbf{R}^\top$ where $\mathbf{R}$ corresponds to a real rectangular matrix from the sparse matrix collection. The multiplicity of λ_1 is $p = 15$ significantly greater than 1. Thus, the potential will be $\sin^2\theta(\mathbf{x}, \mathbf{V}_p) = 1 - \|\mathbf{V}_p^\top \mathbf{x}\|_2^2$ for $\mathbf{x}$ satisfying $\|\mathbf{x}\|_2 = 1$, instead of $\sin^2\theta(\mathbf{x}, \mathbf{v}_1)$. The advantages of our SI-rgEIGS are observed again and the SI-rgEIGS (F.A.) works even better, compared to others except for the Lanczos algorithm. On this special dataset, Lanczos achieves the best performance in terms of the running time. This is because the solution space of this low-rank dataset is relatively much larger than those of others such that the Lanczos with

the minimum memory consumption can also work well[8]. However, it is worth noting that it still consumes more memory than other algorithms.

6.2.2 Generalized Case

We use two challenging real datasets[9] for generalized eigenvalue problems. The challenges spring from the clustered generalized eigenvalues which often result in small values of relative eigengap. Statistics of the data are given in Table 2. We compare our SI-rgGenEIGS (i.e., Algorithm 4) and SI-rgGenEIGS (F.A.) (i.e., Algorithm 5) with the SI-PM-gen (Wang et al., 2016). The shift parameter now is obtained using Algorithm 6 with parameter $m = 18$. The BB step-sizes are used as well. As shown in Figure 8, our SI-rgGenEIGS and SI-rgGenEIGS (F.A.) can work well while the SI-PM-gen fails.

Table 2: Statistics of the matrix pair data.

Matrix pair	n	nnz($\mathbf{A}$)	nnz($\mathbf{B}$)
Lapla3	5795	136565	141779
Lapla5	18903	455337	489875

Table 3: Statistics of the CCA data

$(\mathbf{X}, \mathbf{Y})$	description	d_x	d_y	n
JW11	acoustic and articulation measurements	273	112	30000
MNIST	left and right halves of images	392	392	60000

We also test the algorithms on two CCA datasets that are given in Table 3. The regularization parameters are set to $r_x = r_y = 0.1$. We compare our Algorithm SI-rgGenEIGS[10] with the SI-PM-gen and CCALin (Ge et al., 2016). The SVRG is used as the least-squares solver. It runs four epochs with each running n iterations as well as the step-size $\eta_x = \frac{1}{\max_i \|\mathbf{X}_i\|_2^2}$ or $\eta_y = \frac{1}{\max_i \|\mathbf{Y}_i\|_2^2}$ in our experiments, where $\mathbf{X}_i$ represents the i-th column of $\mathbf{X}$. The shift parameter is obtained using Algorithm 8 for which the parameters are set as follows: $\tilde{\Delta} = 0.06$ (following Wang et al. (2016)), $m_1 = 2$, and $\tilde{\epsilon} = \frac{1}{3084}(\frac{\tilde{\Delta}}{18})^{m_1-1}$. We also use the output of Algorithm 8 to warm-start the three algorithms. The SI-rgGenEIS adopts the BB step-size scheme with initial step-size $\alpha_0 = 10^{-2}$. The following three quality measures are used:

$$\sin^2\theta(\mathbf{x}_t, \frac{1}{\sqrt{2}}\begin{pmatrix}\phi_1\\ \psi_1\end{pmatrix}),\ \sin^2\theta(\tilde{\mathbf{x}}_t^{(1)}, \phi_1),\ \text{and}\ \sin^2\theta(\tilde{\mathbf{x}}_t^{(2)}, \psi_1),$$

8. To understand this rare case, we need to know how Lanczos essentially works. The Lanczos algorithm constructs and repeatedly augments the Krylov subspace (Golub and Van Loan, 2013; Watkins, 2007) which, further by periodical restarting, becomes increasingly close to top invariant subspaces of the given matrix. The solution exists in the first sought invariant subspace and can be recovered by the Rayleigh-Rits procedure on this subspace. Due to the low rank of the dataset and a large dimensionality of the solution space, the demand on the memory consumption becomes lower accordingly.
9. `http://faculty.smu.edu/yzhou/data/matrices.htm`
10. Note that the postprocessing step (30) is required.

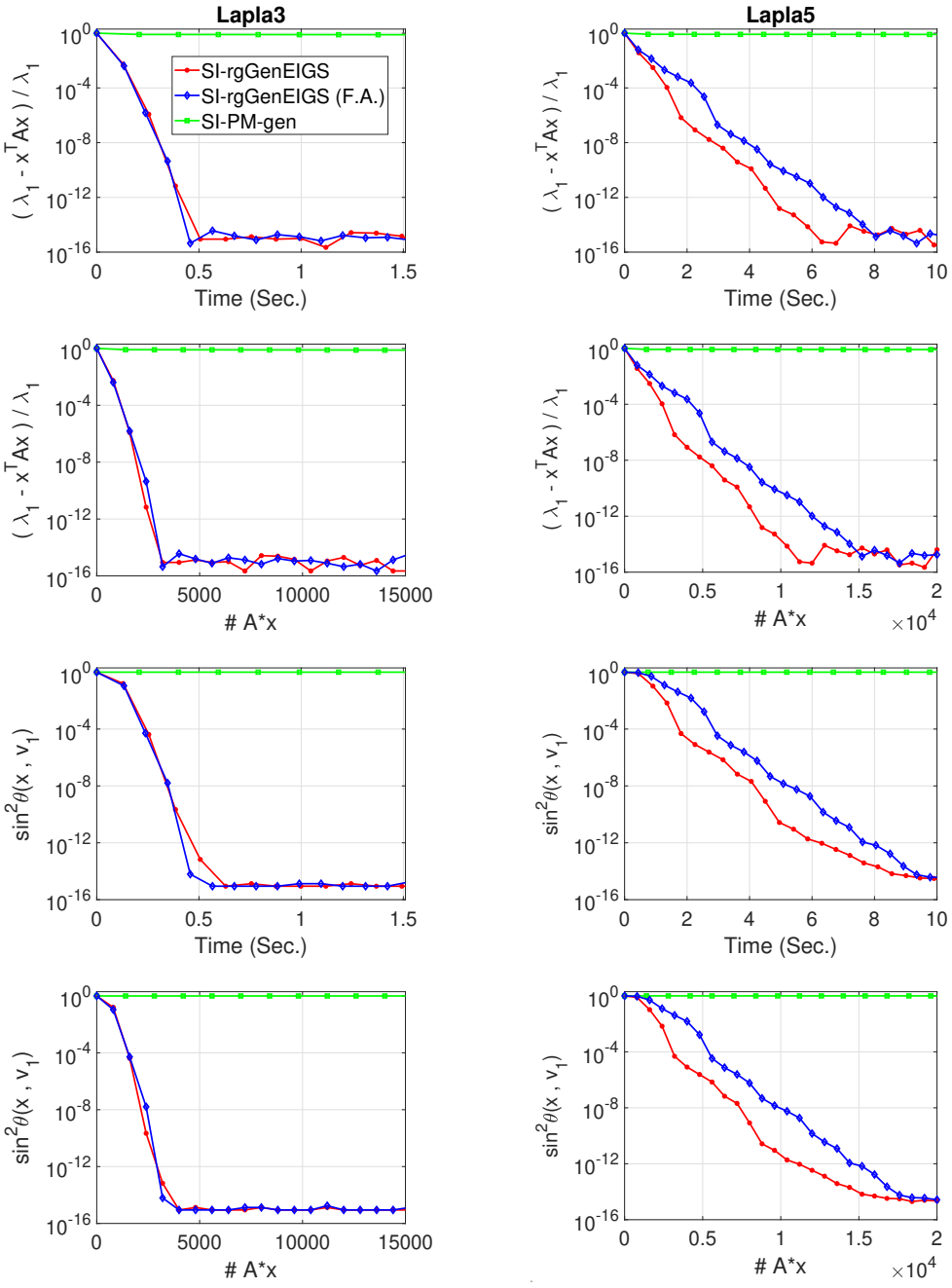

Figure 8: Algorithms for generalized eigenvalue problems on real data.

where $\tilde{\mathbf{x}}_t^{(1)}$ and $\tilde{\mathbf{x}}_t^{(2)}$ are defined similarly to Equation (30) and calculated during iterations as is done at the postprocessing step for evaluation purpose. For brevity, denote

$$\theta \triangleq \theta(\mathbf{x}_t, \frac{1}{\sqrt{2}} \begin{pmatrix} \phi_1 \\ \psi_1 \end{pmatrix}),\ \theta_\phi \triangleq \theta(\tilde{\mathbf{x}}_t^{(1)}, \phi_1),\ \text{and}\ \theta_\psi \triangleq \theta(\tilde{\mathbf{x}}_t^{(2)}, \phi_1).$$

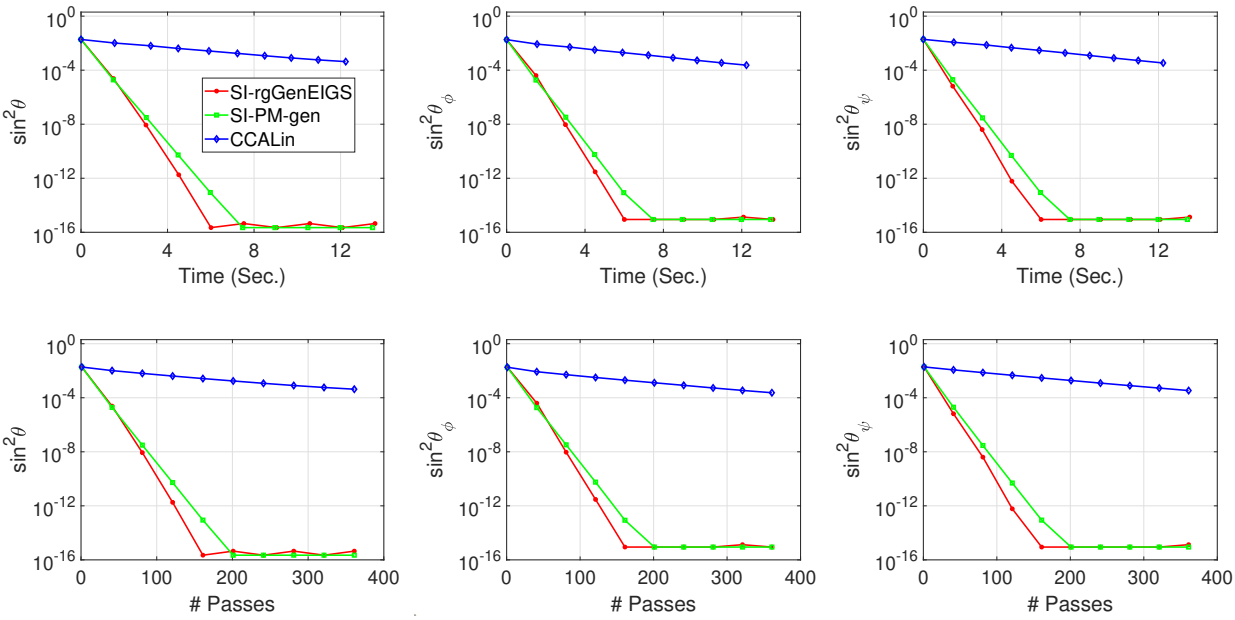

Figure 9: Algorithms for generalized eigenvalue problems on CCA data JW11.

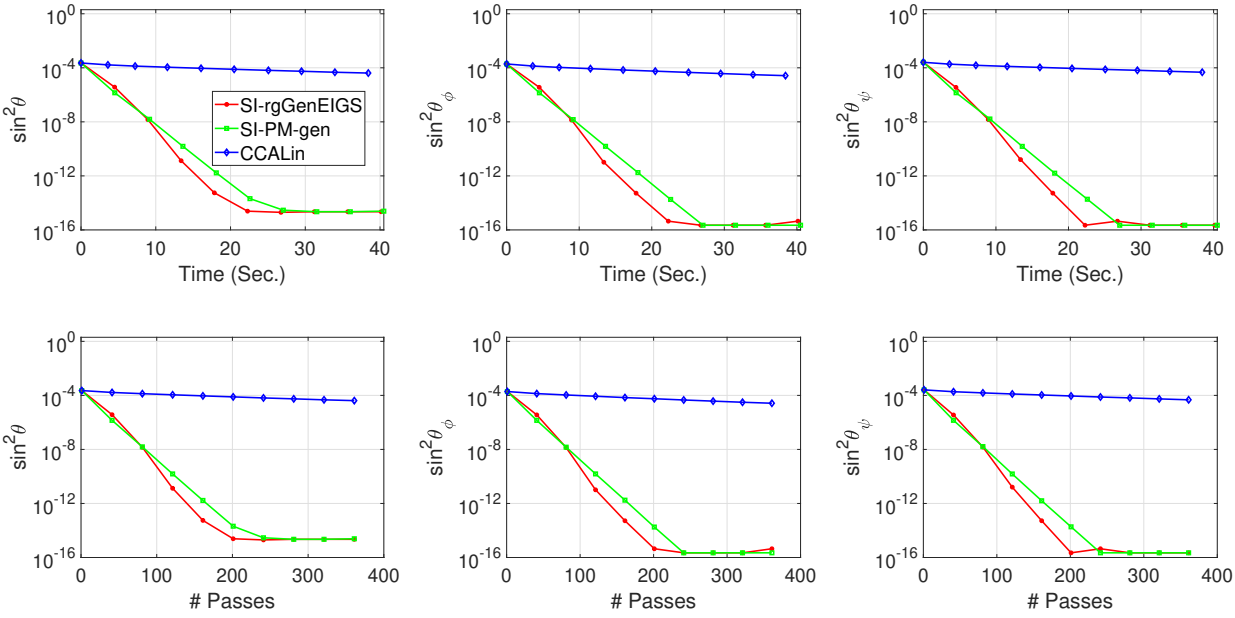

Figure 10: Algorithms for generalized eigenvalue problems on CCA data MNIST.

We plot the convergence curves of three algorithms in terms of the wall-clock time and the number of passes over data for each measure in Figure 9 and Figure 10, where the wart-starting time is excluded. As we can see, both SI-rgGenEIGS and SI-PM-gen achieve much

faster convergence compared to the CCALin which runs inexact power iterations without preconditioning. Furthermore, our SI-rgGenEIGS converges faster than the SI-PM-gen on both datasets in terms of each measure.

7. Conclusion

This paper proposes the first (Riemannian) search method for eigenvector computation with an optimal convergence rate. In order to achieve this rate, the shift-and-invert preconditioning is incorporated into the first-order Riemannian optimization framework and gives rise to inexact Riemannian gradients. Compared to previous convergence rates of search methods, it attains a quadratic improvement in terms of gap dependence. The novelty of our analysis lies in dissecting an equivalent form of the problem where the Riemannian metric is given by the shifted matrix. The analysis is extended to the problem of generalized eigenvector computation as well. Empirically, we demonstrate that the proposed search method can significantly outperform the state-of-the-art projection methods by leveraging the privilege of search methods, that is, different step-size schemes. For future work, it is also interesting to investigate other techniques for accelerating search methods for our problem.

References

Pierre-Antoine Absil, Robert E. Mahony, and Rodolphe Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, 2008.

Kwangjun Ahn and Suvrit Sra. From nesterov's estimate sequence to riemannian acceleration. In *Proceedings of the Conference on Learning Theory (COLT)*, pages 84–118, Virtual Event [Graz, Austria], 2020.

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

Zeyuan Allen-Zhu and Yuanzhi Li. Doubly accelerated methods for faster CCA and generalized eigendecomposition. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pages 98–106, Sydney, Australia, 2017.

Raman Arora, Andrew Cotter, and Nati Srebro. Stochastic optimization of PCA with capped MSG. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1815–1823, Lake Tahoe, NV, 2013.

Maria-Florina Balcan, Simon Shaolei Du, Yining Wang, and Adams Wei Yu. An improved gap-dependency analysis of the noisy power method. In *Proceedings of the 29th Conference on Learning Theory (COLT)*, pages 284–309, New York, 2016.

Akshay Balsubramani, Sanjoy Dasgupta, and Yoav Freund. The fast convergence of incremental PCA. In *Advances in Neural Information Processing Systems (NIPS)*, pages 3174–3182, Lake Tahoe, NV, 2013.

Jonathan Barzilai and Jonathan M Borwein. Two-point step size gradient methods. *IMA Journal of Numerical Analysis*, 8:141–148, 01 1988.

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

Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends in Machine Learning*, 8(3-4):231–357, 2015.

Jianqing Fan, Qiang Sun, Wen-Xin Zhou, and Ziwei Zhu. Principal component analysis for big data. *arXiv preprint arXiv:1801.01602*, 2018.

Dan Garber and Elad Hazan. Fast and simple pca via convex optimization. *arXiv preprint arXiv:1509.05647*, 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 *Proceedings of the 33nd International Conference on Machine Learning (ICML)*, pages 2626–2634, New York City, NY, 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 *Proceedings of the 33nd International Conference on Machine Learning (ICML)*, pages 2741–2750, New York City, NY, 2016.

Gene H Golub and Charles F Van Loan. *Matrix Computations, forth edition*. Johns Hopkins University Press, 2013.

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

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

Trevor Hastie, Rahul Mazumder, Jason D. Lee, and Reza Zadeh. Matrix completion and low-rank SVD via fast alternating least squares. *J. Mach. Learn. Res.*, 16:3367–3402, 2015.

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

Harold Hotelling. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377, December 1936.

Bruno Iannazzo and Margherita Porcelli. The riemannian barzilai–borwein method with nonmonotone line search and the matrix geometric mean computation. *IMA Journal of Numerical Analysis*, 38(1):495–517, 2017.

Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems (NIPS)*, pages 315–323, Lake Tahoe, NV, 2013.

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 (NIPS)*, pages 2056–2064, Barcelona, Spain, 2016.

Hongzhou Lin, Julien Mairal, and Zaïd Harchaoui. A universal catalyst for first-order optimization. In *Advances in Neural Information Processing Systems (NIPS)*, pages 3384–3392, Montreal, Canada, 2015.

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 *Proceedings of the 33nd International Conference on Machine Learning (ICML)*, pages 1158–1167, New York City, NY, 2016.

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 *Advances in Neural Information Processing Systems (NIPS)*, pages 1396–1404, Montreal, Canada, 2015.

Yurii Nesterov and Sebastian U. Stich. Efficiency of the accelerated coordinate descent method on structured optimization problems. *SIAM Journal on Optimization*, 27(1):110–123, 2017. doi: 10.1137/16M1060182. URL `https://doi.org/10.1137/16M1060182`.

Yurii E. Nesterov. *Introductory Lectures on Convex Optimization - A Basic Course*, volume 87 of *Applied Optimization*. Springer, 2004.

Yurii E. Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM J. Optim.*, 22(2):341–362, 2012.

Andrew Y. Ng, Michael I. Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems (NIPS)*, pages 849–856, Vancouver, Canada, 2001.

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

Yousef Saad. *Numerical methods for large eigenvalue problems: revised edition*. SIAM, 2011.

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

Ohad Shamir. Fast stochastic algorithms for SVD and PCA: convergence properties and convexity. In *Proceedings of the 33nd International Conference on Machine Learning (ICML)*, pages 248–256, New York City, NY, 2016a.

Ohad Shamir. Convergence of stochastic gradient descent for PCA. In *Proceedings of the 33nd International Conference on Machine Learning (ICML)*, pages 257–265, New York City, NY, 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. In *Proceedings of the Conference on Algorithmic Learning Theory (ALT)*, pages 806–820, Lanzarote, Canary Islands, Spain, 2018.

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 (NIPS)*, pages 766–774, Barcelona, Spain, 2016.

David S. Watkins. *The Matrix Eigenvalue Problem: GR and Krylov Subspace Methods*. Society for Industrial and Applied Mathematics, 2007.

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

James H. Wilkinson. *The Algebraic Eigenvalue Problem*. Clarendon Press, 1988.

Peng Xu, Bryan D. He, Christopher De Sa, Ioannis Mitliagkas, and Christopher Ré. Accelerated stochastic power iteration. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 58–67, Playa Blanca, Lanzarote, Canary Islands, Spain, 2018a.

Zhiqiang Xu and Ping Li. Towards practical alternating least-squares for CCA. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 14737–14746, Vancouver, Canada, 2019.

Zhiqiang Xu and Ping Li. On the faster alternating least-squares for CCA. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 1621–1629, Virtual Event, 2021a.

Zhiqiang Xu and Ping Li. A comprehensively tight analysis of gradient descent for pca. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021b.

Zhiqiang Xu, Yiping Ke, and Xin Gao. A fast algorithm for matrix eigen-decompositionn. In *Proceedings of the Thirty-Third Conference on Uncertainty in Artificial Intelligence (UAI)*, Sydney, Australia, 2017.

Zhiqiang Xu, Xin Cao, and Xin Gao. Convergence analysis of gradient descent for eigenvector computation. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI)*, pages 2933–2939, Stockholm, Sweden, 2018b.

Hongyi Zhang and Suvrit Sra. An estimate sequence for geodesically convex optimization. In *Proceedings of the Conference On Learning Theory (COLT)*, pages 1703–1723, Stockholm, Sweden, 2018.

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

Zhihua Zhang. The singular value decomposition, applications and beyond. *arXiv preprint arXiv:1510.08532*, 2015.

Yunkai Zhou, Yousef Saad, Murilo L. Tiago, and James R. Chelikowsky. Self-consistent-field calculations using chebyshev-filtered subspace iteration. *J. Comput. Physics*, 219 (1):172–184, 2006.

Appendix

Algorithm 7 (Garber and Hazan, 2015; Wang et al., 2018) locate $\sigma = \lambda_1 + c\Delta_p$

Input: matrix $\mathbf{A}$ and lower estimate η satisfying $c_1\Delta_p \leq \eta \leq c_2\Delta_p$ where $0 < c_1 < c_2 \leq 1$, least-squares solver $\text{ls}(\tilde{\mathbf{A}}, \tilde{\mathbf{b}}, \tilde{\mathbf{x}}_0)$.
$\tilde{\mathbf{a}}_0 = \frac{\mathbf{r}}{\|\mathbf{r}\|_2}$ where $\mathbf{r} \in \mathbb{R}^{n\times 1}$ and $r_i \sim \mathcal{N}(0,1)$
$s = 0$ and $\sigma_s = 1 + \eta$
repeat
  $\mathbf{a}_0 = \tilde{\mathbf{a}}_s$
  for $t = 1, 2, \cdots, m$ **do**
    $\hat{\mathbf{a}}_t \approx \text{ls}(\sigma_s\mathbf{I} - \mathbf{A}, \mathbf{a}_{t-1}, \frac{\mathbf{a}_{t-1}}{\mathbf{a}_{t-1}^\top(\sigma_s\mathbf{I}-\mathbf{A})\mathbf{a}_{t-1}})$ and $\mathbf{a}_t = \frac{\hat{\mathbf{a}}_t}{\|\hat{\mathbf{a}}_t\|_2}$
  end for
  $\tilde{\mathbf{a}}_{s+1} = \mathbf{a}_m$
  $\mathbf{w} \approx \text{ls}(\sigma_s\mathbf{I} - \mathbf{A}, \tilde{\mathbf{a}}_{s+1}, \frac{\tilde{\mathbf{a}}_{s+1}}{\tilde{\mathbf{a}}_{s+1}^\top(\sigma_s\mathbf{I}-\mathbf{A})\tilde{\mathbf{a}}_{s+1}})$
  $\eta_{s+1} = \frac{1}{2}\frac{1}{\tilde{\mathbf{a}}_{s+1}^\top\mathbf{w} - \frac{1}{8}(1+\frac{1-c_2}{c_2}\eta)}$ and $\sigma_{s+1} = \sigma_s - \frac{1}{2}\eta_{s+1}$
  $s \leftarrow s + 1$
until $\eta_s \leq \eta$
Output: $\sigma = \sigma_s$ and $\mathbf{x}_0 = \tilde{\mathbf{a}}_s$

Algorithm 8 (Wang et al., 2016) locate $\sigma = \lambda_1 + c\Delta_p$

Input: matrix pair $(\mathbf{A}, \mathbf{B})$ and lower estimate η satisfying $c_1\Delta_p \leq \eta \leq c_2\Delta_p$ where $0 < c_1 < c_2 \leq 1$, least-squares solver $\text{ls}(\tilde{\mathbf{A}}, \tilde{\mathbf{b}}, \tilde{\mathbf{x}}_0)$.
$\tilde{\mathbf{a}}_0 = \frac{\mathbf{r}}{\|\mathbf{r}\|_\mathbf{B}}$ where $\mathbf{r} \in \mathbb{R}^{n\times 1}$ and $r_i \sim \mathcal{N}(0,1)$
$s = 0$ and $\sigma_s = 1 + \eta$
repeat
  $\mathbf{a}_0 = \tilde{\mathbf{a}}_s$
  for $t = 1, 2, \cdots, m$ **do**
    $\hat{\mathbf{a}}_t \approx \text{ls}(\sigma_s\mathbf{B} - \mathbf{A}, \mathbf{B}\mathbf{a}_{t-1}, \frac{\mathbf{a}_{t-1}^\top\mathbf{B}\mathbf{a}_{t-1}}{\mathbf{a}_{t-1}^\top(\sigma_s\mathbf{B}-\mathbf{A})\mathbf{a}_{t-1}}\mathbf{a}_{t-1})$ such that $l_t(\hat{\mathbf{a}}_t) \leq \min l_t(\mathbf{x}) + \tilde{\epsilon}$
    $\mathbf{a}_t = \frac{\hat{\mathbf{a}}_t}{\|\hat{\mathbf{a}}_t\|_2}$
  end for
  $\tilde{\mathbf{a}}_{s+1} = \mathbf{a}_m$
  $\mathbf{w} \approx \text{ls}(\sigma_s\mathbf{B} - \mathbf{A}, \mathbf{B}\tilde{\mathbf{a}}_{s+1}, \frac{\tilde{\mathbf{a}}_{s+1}^\top\mathbf{B}\tilde{\mathbf{a}}_{s+1}}{\tilde{\mathbf{a}}_{s+1}^\top(\sigma_s\mathbf{B}-\mathbf{A})\tilde{\mathbf{a}}_{s+1}}\tilde{\mathbf{a}}_{s+1})$ such that $l_s(\mathbf{w}) \leq \min l_s(\mathbf{x}) + \tilde{\epsilon}$
  $\eta_{s+1} = \frac{1}{\tilde{\mathbf{a}}_{s+1}^\top\mathbf{B}\mathbf{w} - 4\sqrt{\tilde{\epsilon}/\tilde{\Delta}}}$ and $\sigma_{s+1} = \sigma_s - \frac{1}{2}\eta_{s+1}$
  $s \leftarrow s + 1$
until $\eta_s \leq \eta$
Output: $\sigma = \sigma_s$ and $\mathbf{x}_0 = \tilde{\mathbf{a}}_s$

Proof of Lemma 3

Note that the full eigenvalue decomposition of $\mathbf{A}$ is

$$\mathbf{A} = \mathbf{V}_q \mathrm{diag}(\lambda_1, \cdots, \lambda_q)\mathbf{V}_q^\top + \mathbf{V}_q^\perp \mathrm{diag}(\lambda_{q+1}, \cdots, \lambda_n)(\mathbf{V}_q^\perp)^\top,$$

where $\mathbf{V}_q^\perp$ represents the orthogonal complement of $\mathbf{V}_q$. We then have that

$$\begin{aligned}
\lambda_1 - \mathbf{x}^\top \mathbf{A}\mathbf{x} &= \lambda_1 - \mathbf{x}^\top \mathbf{V}_q \mathrm{diag}(\lambda_1, \cdots, \lambda_q)\mathbf{V}_q^\top \mathbf{x} - \mathbf{x}^\top \mathbf{V}_q^\perp \mathrm{diag}(\lambda_{q+1}, \cdots, \lambda_n)(\mathbf{V}_q^\perp)^\top \mathbf{x} \\
&\geq \lambda_1 - \lambda_1 \mathbf{x}^\top \mathbf{V}_q \mathbf{V}_q^\top \mathbf{x} - \mathbf{x}^\top \mathbf{V}_q^\perp \mathrm{diag}(\lambda_{q+1}, \cdots, \lambda_n)(\mathbf{V}_q^\perp)^\top \mathbf{x} \\
&\geq \lambda_1 \sin^2\theta(\mathbf{x}, \mathbf{V}_q) - \lambda_{q+1} \mathbf{x}^\top \mathbf{V}_q^\perp (\mathbf{V}_q^\perp)^\top \mathbf{x} \\
&= \lambda_1 \sin^2\theta(\mathbf{x}, \mathbf{V}_q) - \lambda_{q+1} \mathbf{x}^\top (\mathbf{I} - \mathbf{V}_q \mathbf{V}_q^\top) \mathbf{x} \\
&= (\lambda_1 - \lambda_{q+1}) \sin^2\theta(\mathbf{x}, \mathbf{V}_q).
\end{aligned}$$

■

Proof of Lemma 4

Noting that $\mathbf{x} = \mathbf{C}^{1/2}\mathbf{y}$, we have

$$\begin{aligned}
\|\tilde{\nabla} h(\mathbf{y})\|_{\mathbf{C}}^2 &= \|\mathbf{C}^{1/2}(\mathbf{I} - \mathbf{y}\mathbf{y}^\top \mathbf{C})\mathbf{C}^{-1}\mathbf{y}\|_2^2 \\
&= \|\mathbf{C}^{1/2}(\mathbf{I} - \mathbf{y}\mathbf{y}^\top \mathbf{C})\mathbf{C}^{-1/2}\mathbf{C}^{-1/2}\mathbf{y}\|_2^2 \\
&= \|(\mathbf{I} - \mathbf{x}\mathbf{x}^\top)\mathbf{C}^{-1/2}\mathbf{y}\|_2^2 = \|(\mathbf{I} - \mathbf{x}\mathbf{x}^\top)\mathbf{C}^{-1}\mathbf{x}\|_2^2.
\end{aligned}$$

For any $\mathbf{v} \in \mathcal{V}_{p,1}$, it holds that

$$\mathbf{C}^{-1} = \mu_1 \mathbf{v}\mathbf{v}^\top + \mathbf{v}_\perp \mathrm{diag}(\mu_2, \cdots, \mu_n)\mathbf{v}_\perp^\top.$$

Plugging in the above equation to the gradient, one gets

$$\begin{aligned}
\|\tilde{\nabla} h(\mathbf{y})\|_{\mathbf{C}}^2 &= \|(\mathbf{I} - \mathbf{x}\mathbf{x}^\top)\mathbf{C}^{-1}\mathbf{x}\|_2^2 \\
&= \|\mathbf{x}_\perp^\top(\mu_1 \mathbf{v}\mathbf{v}^\top + \mathbf{v}_\perp \mathrm{diag}(\mu_2, \cdots, \mu_n)\mathbf{v}_\perp^\top)\mathbf{x}\|_2^2 \\
&\leq 2\mu_1^2 \|\mathbf{x}_\perp^\top \mathbf{v}\|^2 + 2\mu_2^2 \|\mathbf{v}_\perp^\top \mathbf{x}\|^2 = 2\mu_1^2(1 - (\mathbf{x}^\top \mathbf{v})^2) + 2\mu_2^2(1 - (\mathbf{v}^\top \mathbf{x})^2) \\
&\leq 4\mu_1^2(1 - (\mathbf{x}^\top \mathbf{v})^2).
\end{aligned}$$

Since the above inequality holds for any $\mathbf{v} \in \mathcal{V}_{p,1}$, we get

$$\begin{aligned}
\|\tilde{\nabla} h(\mathbf{y})\|_{\mathbf{C}}^2 &\leq 4\mu_1^2 \min_{\mathbf{v} \in \mathcal{V}_{p,1}} (1 - (\mathbf{x}^\top \mathbf{v})^2) = 4\mu_1^2 \sin^2\theta(\mathbf{x}, \mathbf{V}_p) \\
&= 4\mu_1^2 \sin^2\theta(\mathbf{C}^{1/2}\mathbf{y}, \mathbf{V}_p).
\end{aligned}$$

■

Proof of Lemma 5

- First note that

$$l_t(\mathbf{C}^{-1}\mathbf{y}_t) = \frac{1}{2}\mathbf{y}_t^\top \mathbf{C}^{-1}\mathbf{C}^{-1}\mathbf{C}^{-1}\mathbf{y}_t - \mathbf{y}_t^\top \mathbf{C}^{-1}\mathbf{y}_t = -\frac{1}{2}\mathbf{y}_t\mathbf{C}^{-1}\mathbf{y}_t.$$

Then we have

$$\begin{aligned}
\frac{1}{2}\|\mathbf{y}-\mathbf{C}^{-1}\mathbf{y}_t\|_{\mathbf{C}}^2 &= \frac{1}{2}(\mathbf{y}-\mathbf{C}^{-1}\mathbf{y}_t)^\top\mathbf{C}(\mathbf{y}-\mathbf{C}^{-1}\mathbf{y}_t)\\
&= \frac{1}{2}\mathbf{y}^\top\mathbf{C}\mathbf{y} - \mathbf{y}_t^\top\mathbf{C}^{-1}\mathbf{C}\mathbf{y} + \frac{1}{2}\mathbf{y}_t^\top\mathbf{C}^{-1}\mathbf{C}\mathbf{C}^{-1}\mathbf{y}_t\\
&= \frac{1}{2}\mathbf{y}^\top\mathbf{C}\mathbf{y} - \mathbf{y}_t^\top\mathbf{y} + \frac{1}{2}\mathbf{y}_t^\top\mathbf{C}^{-1}\mathbf{y}_t\\
&= l_t(\mathbf{y}) - l_t(\mathbf{C}^{-1}\mathbf{y}_t) = \epsilon_t(\mathbf{y}).
\end{aligned}$$

We next show that $2\epsilon_t(\|\mathbf{y}_t\|_2^2\mathbf{y}_t) \leq \mu_1^2\sin^2\theta(\mathbf{C}^{1/2}\mathbf{y}_t, \mathbf{V}_p)$. Let $r(\gamma) = \epsilon_t(\gamma\mathbf{y}_t)$. Noting that $\mathbf{y}_t \in \mathrm{St}_{\mathbf{C}}(n,1)$, we have

$$\begin{aligned}
r(\gamma) &= l_t(\gamma\mathbf{y}_t) - l_t(\mathbf{C}^{-1}\mathbf{y}_t)\\
&= \frac{\gamma^2}{2}\mathbf{y}_t^\top\mathbf{C}\mathbf{y}_t - \gamma\mathbf{y}_t^\top\mathbf{y}_t - l_t(\mathbf{C}^{-1}\mathbf{y}_t)\\
&= \frac{\gamma^2}{2} - \gamma\|\mathbf{y}_t\|_2^2 - l_t(\mathbf{C}^{-1}\mathbf{y}_t),
\end{aligned}$$

which is minimized at the root of $r'(\gamma) = 0$, i.e., $\gamma = \|\mathbf{y}_t\|_2^2$. Thus, we have

$$\begin{aligned}
\epsilon_t(\|\mathbf{y}_t\|_2^2\mathbf{y}_t) &\leq l_t(\mu_1\mathbf{y}_t) - l_t(\mathbf{C}^{-1}\mathbf{y}_t)\\
&= \frac{\mu_1^2}{2}\mathbf{y}_t^\top\mathbf{C}\mathbf{y}_t - \mu_1\mathbf{y}_t^\top\mathbf{y}_t + \frac{1}{2}\mathbf{y}_t^\top\mathbf{C}^{-1}\mathbf{y}_t\\
&= \frac{\mu_1^2}{2}\sum_{i=1}^n\frac{(\mathbf{v}_i^\top\mathbf{y}_t)^2}{\mu_i} - \mu_1\sum_{i=1}^n(\mathbf{v}_i^\top\mathbf{y}_t)^2 + \frac{1}{2}\sum_{i=1}^n\mu_i(\mathbf{v}_i^\top\mathbf{y}_t)^2\\
&= \sum_{i=1}^n(\frac{\mu_1^2}{2\mu_i} - \mu_1 + \frac{1}{2}\mu_i)(\mathbf{v}_i^\top\mathbf{y}_t)^2 = \sum_{i=1}^n\frac{(\mu_1-\mu_i)^2}{2\mu_i}(\mathbf{v}_i^\top\mathbf{y}_t)^2\\
&= \sum_{i=1}^n\frac{(\mu_1-\mu_i)^2}{2\mu_i}(\mathbf{v}_i^\top\mathbf{C}^{-1/2}\mathbf{C}^{1/2}\mathbf{y}_t)^2 = \sum_{i=1}^n\frac{(\mu_1-\mu_i)^2}{2\mu_i}(\sqrt{\mu_i}\mathbf{v}_i^\top\mathbf{C}^{1/2}\mathbf{y}_t)^2\\
&= \frac{1}{2}\sum_{i=1}^n(\mu_1-\mu_i)^2(\mathbf{v}_i^\top\mathbf{C}^{1/2}\mathbf{y}_t)^2 = \frac{1}{2}\sum_{i>p}^n(\mu_1-\mu_i)^2(\mathbf{v}_i^\top\mathbf{C}^{1/2}\mathbf{y}_t)^2\\
&\leq \frac{\mu_1^2}{2}\sum_{i=p+1}^n(\mathbf{v}_i^\top\mathbf{C}^{1/2}\mathbf{y}_t)^2 = \frac{\mu_1^2}{2}(1 - \sum_{i=1}^p(\mathbf{v}_i^\top\mathbf{C}^{1/2}\mathbf{y}_t)^2)\\
&= \frac{\mu_1^2}{2}\sin^2\theta(\mathbf{C}^{1/2}\mathbf{y}_t, \mathbf{V}_p).
\end{aligned}$$

- The complexity can be obtained by noting that the Hessian of $l_t(\mathbf{y})$ satisfies

$$\frac{1}{\mu_1}\mathbf{I} \preccurlyeq \text{Hessian}(l_t(\mathbf{y})) = \mathbf{C} \preccurlyeq \frac{1}{\mu_n}\mathbf{I}.$$

That is, $l_t(\mathbf{y})$ is $\frac{1}{\mu_1}$-strongly convex and $\frac{1}{\mu_n}$-smooth. Thus, Nesterov's accelerated gradient descent (Nesterov, 2004; Bubeck, 2015) takes

$$O(\sqrt{\frac{\frac{1}{\mu_n}}{\frac{1}{\mu_1}}}\log\frac{l_t(\|\mathbf{y}_t\|_2^2\mathbf{y}_t) - l_t(\mathbf{C}^{-1}\mathbf{y}_t)}{\epsilon_t(\widehat{\mathbf{C}^{-1}\mathbf{y}_t})}) = O(\sqrt{\frac{\lambda_1}{\Delta_\epsilon}}\log\frac{\epsilon_t(\|\mathbf{y}_t\|_2^2\mathbf{y}_t)}{\epsilon_t(\widehat{\mathbf{C}^{-1}\mathbf{y}_t})})$$

complexity for the $\frac{1}{\mu_1}$-strongly convex and $\frac{1}{\mu_n}$-smooth function $l_t(\mathbf{y})$ and

$$O(\|\|\mathbf{y}_t\|_2^2\mathbf{y}_t - \mathbf{C}^{-1}\mathbf{y}_t\|_2\sqrt{\frac{\frac{2}{\mu_n}}{\epsilon_t(\widehat{\mathbf{C}^{-1}\mathbf{y}_t})}}) = O(\|(\|\mathbf{y}_t\|_2^2\mathbf{I} - \mathbf{C}^{-1})\mathbf{y}_t\|_2\sqrt{\frac{\lambda_1}{\epsilon_t(\widehat{\mathbf{C}^{-1}\mathbf{y}_t})}})$$

complexity for the convex and $\frac{1}{\mu_n}$-smooth function $l_t(\mathbf{y})$ to reach sub-optimality $\epsilon_t(\widehat{\mathbf{C}^{-1}\mathbf{y}_t})$, where we have used that

$$\begin{aligned}\frac{1}{\mu_1} &= \sigma - \lambda_1 = \lambda_1 + c\Delta_p - \lambda_1 = c\Delta_p,\\ \frac{1}{\mu_n} &= \sigma - \lambda_n \le \sigma = \lambda_1 + c(\lambda_1 - \lambda_{p+1}) \le (1+c)\lambda_1.\end{aligned}$$

■

Proof of Lemma 8

Note that the full generalized eigenvalue decomposition of $(\mathbf{A}, \mathbf{B})$ is

$$\mathbf{A} = \mathbf{B}(\mathbf{V}_q\text{diag}(\lambda_1, \cdots, \lambda_q)\mathbf{V}_q^\top + \mathbf{V}_q^\perp\text{diag}(\lambda_{q+1}, \cdots, \lambda_n)(\mathbf{V}_q^\perp)^\top)\mathbf{B}, \tag{31}$$

where $\mathbf{V}_q^\perp$ represents the orthogonal complement of $\mathbf{V}_q$ in the $\mathbf{B}$-norm, i.e., $\mathbf{V}_q^\top\mathbf{B}\mathbf{V}_q^\perp = \mathbf{0}$. We then have that

$$\begin{aligned}\lambda_1 - \mathbf{x}^\top\mathbf{A}\mathbf{x} &= \lambda_1 - \mathbf{x}^\top\mathbf{B}\mathbf{V}_q\text{diag}(\lambda_1, \cdots, \lambda_q)\mathbf{V}_q^\top\mathbf{B}\mathbf{x}\\ &\quad -\mathbf{x}^\top\mathbf{B}\mathbf{V}_q^\perp\text{diag}(\lambda_{q+1}, \cdots, \lambda_n)(\mathbf{V}_q^\perp)^\top\mathbf{B}\mathbf{x}\\ &\ge \lambda_1 - \lambda_1\mathbf{x}^\top\mathbf{B}\mathbf{V}_q\mathbf{V}_q^\top\mathbf{B}\mathbf{x} - \mathbf{x}^\top\mathbf{B}\mathbf{V}_q^\perp\text{diag}(\lambda_{q+1}, \cdots, \lambda_n)(\mathbf{V}_q^\perp)^\top\mathbf{B}\mathbf{x}\\ &\ge \lambda_1\sin^2\theta(\mathbf{x}, \mathbf{V}_q) - \lambda_{q+1}\mathbf{x}^\top\mathbf{B}\mathbf{V}_q^\perp(\mathbf{V}_q^\perp)^\top\mathbf{B}\mathbf{x}\\ &= \lambda_1\sin^2\theta(\mathbf{x}, \mathbf{V}_q) - \lambda_{q+1}\mathbf{x}^\top\mathbf{B}^{1/2}(\mathbf{I} - \mathbf{B}^{1/2}\mathbf{V}_q\mathbf{V}_q^\top\mathbf{B}^{1/2})\mathbf{B}^{1/2}\mathbf{x}\\ &= (\lambda_1 - \lambda_{q+1})\sin^2\theta(\mathbf{x}, \mathbf{V}_q).\end{aligned}$$

■

Proof of Lemma 9

Noting in Equation (29) that for $\mathbf{x} \in \mathrm{St}_{\mathbf{B}}(n,1)$ the relation

$$\mathbf{y} = \mathbf{B}^{-1/2}\mathbf{C}^{-1/2}\mathbf{B}^{1/2}\mathbf{x} \in \mathrm{St}_{\sigma\mathbf{B}-\mathbf{A}}(n,1)$$

holds, we have

$$\begin{aligned}
& \|\tilde{\nabla} h(\mathbf{y})\|^2_{\sigma\mathbf{B}-\mathbf{A}} \\
= \; & \|((\sigma\mathbf{B}-\mathbf{A})^{-1} - \mathbf{y}\mathbf{y}^\top)\mathbf{B}\mathbf{y}\|^2_{\sigma\mathbf{B}-\mathbf{A}} \\
= \; & \mathbf{y}^\top\mathbf{B}((\sigma\mathbf{B}-\mathbf{A})^{-1} - \mathbf{y}\mathbf{y}^\top)^\top(\sigma\mathbf{B}-\mathbf{A})((\sigma\mathbf{B}-\mathbf{A})^{-1} - \mathbf{y}\mathbf{y}^\top)\mathbf{B}\mathbf{y} \\
= \; & \mathbf{y}^\top\mathbf{B}((\sigma\mathbf{B}-\mathbf{A})^{-1} - \mathbf{y}\mathbf{y}^\top - \mathbf{y}\mathbf{y}^\top + \mathbf{y}\mathbf{y}^\top(\sigma\mathbf{B}-\mathbf{A})\mathbf{y}\mathbf{y}^\top)\mathbf{B}\mathbf{y} \\
= \; & \mathbf{y}^\top\mathbf{B}((\sigma\mathbf{B}-\mathbf{A})^{-1} - \mathbf{y}\mathbf{y}^\top)\mathbf{B}\mathbf{y} \\
= \; & \mathbf{x}^\top\mathbf{B}^{1/2}\mathbf{C}^{-\frac{1}{2}}\mathbf{B}^{-\frac{1}{2}}\mathbf{B}(\mathbf{B}^{-\frac{1}{2}}\mathbf{C}^{-1}\mathbf{B}^{-\frac{1}{2}} - \mathbf{B}^{-\frac{1}{2}}\mathbf{C}^{-\frac{1}{2}}\mathbf{B}^{\frac{1}{2}}\mathbf{x}\mathbf{x}^\top\mathbf{B}^{\frac{1}{2}}\mathbf{C}^{-\frac{1}{2}}\mathbf{B}^{-\frac{1}{2}})\mathbf{B}\mathbf{B}^{-\frac{1}{2}}\mathbf{C}^{-\frac{1}{2}}\mathbf{B}^{\frac{1}{2}}\mathbf{x} \\
= \; & \mathbf{x}^\top\mathbf{B}^{1/2}\mathbf{C}^{-\frac{1}{2}}(\mathbf{C}^{-1} - \mathbf{C}^{-\frac{1}{2}}\mathbf{B}^{\frac{1}{2}}\mathbf{x}\mathbf{x}^\top\mathbf{B}^{\frac{1}{2}}\mathbf{C}^{-\frac{1}{2}})\mathbf{C}^{-\frac{1}{2}}\mathbf{B}^{\frac{1}{2}}\mathbf{x} \\
= \; & \mathbf{x}^\top\mathbf{B}^{1/2}\mathbf{C}^{-1}(\mathbf{I} - \mathbf{B}^{\frac{1}{2}}\mathbf{x}\mathbf{x}^\top\mathbf{B}^{\frac{1}{2}})\mathbf{C}^{-1}\mathbf{B}^{\frac{1}{2}}\mathbf{x} \\
= \; & \|(\mathbf{I} - \mathbf{B}^{\frac{1}{2}}\mathbf{x}\mathbf{x}^\top\mathbf{B}^{\frac{1}{2}})\mathbf{C}^{-1}\mathbf{B}^{\frac{1}{2}}\mathbf{x}\|_2^2.
\end{aligned}$$

For any $\mathbf{v} \in \mathcal{V}_{p,1}$, by Equation (31) it holds that

$$\begin{aligned}
\mathbf{C}^{-1} &= (\sigma\mathbf{I} - \mathbf{B}^{-\frac{1}{2}}\mathbf{A}\mathbf{B}^{-\frac{1}{2}})^{-1} \\
&= \mu_1\mathbf{B}^{\frac{1}{2}}\mathbf{v}\mathbf{v}^\top\mathbf{B}^{\frac{1}{2}} + \mathbf{B}^{\frac{1}{2}}\mathbf{v}_\perp\mathrm{diag}(\mu_2,\cdots,\mu_n)\mathbf{v}_\perp^\top\mathbf{B}^{\frac{1}{2}}. \qquad (32)
\end{aligned}$$

Plugging in the above equation to the gradient, one gets

$$\begin{aligned}
\|\tilde{\nabla} h(\mathbf{y})\|^2_{\sigma\mathbf{B}-\mathbf{A}} &= \|(\mathbf{I} - \mathbf{B}^{\frac{1}{2}}\mathbf{x}\mathbf{x}^\top\mathbf{B}^{\frac{1}{2}})\mathbf{C}^{-1}\mathbf{B}^{\frac{1}{2}}\mathbf{x}\|_2^2 \\
&= \|\mathbf{x}_\perp^\top\mathbf{B}^{\frac{1}{2}}(\mu_1\mathbf{B}^{\frac{1}{2}}\mathbf{v}\mathbf{v}^\top\mathbf{B}^{\frac{1}{2}} + \mathbf{B}^{\frac{1}{2}}\mathbf{v}_\perp\mathrm{diag}(\mu_2,\cdots,\mu_n)\mathbf{v}_\perp^\top\mathbf{B}^{\frac{1}{2}})\mathbf{B}^{\frac{1}{2}}\mathbf{x}\|_2^2 \\
&\leq 2\mu_1^2\|\mathbf{x}_\perp^\top\mathbf{B}\mathbf{v}\|^2 + 2\mu_2^2\|\mathbf{v}_\perp^\top\mathbf{B}\mathbf{x}\|^2 \\
&= 2\mu_1^2(1 - (\mathbf{x}^\top\mathbf{B}\mathbf{v})^2) + 2\mu_2^2(1 - (\mathbf{v}^\top\mathbf{B}\mathbf{x})^2) \\
&\leq 4\mu_1^2(1 - (\mathbf{x}^\top\mathbf{B}\mathbf{v})^2).
\end{aligned}$$

Since the above inequality holds for any $\mathbf{v} \in \mathcal{V}_{p,1}$, we get that

$$\begin{aligned}
\|\tilde{\nabla} h(\mathbf{y})\|^2_{\sigma\mathbf{B}-\mathbf{A}} &\leq 4\mu_1^2 \min_{\mathbf{v}\in\mathcal{V}_{p,1}} (1 - (\mathbf{x}^\top\mathbf{B}\mathbf{v})^2) = 4\mu_1^2\sin^2\theta(\mathbf{x}, \mathbf{V}_p) \\
&= 4\mu_1^2\sin^2\theta(\mathbf{B}^{-1/2}\mathbf{C}^{1/2}\mathbf{B}^{1/2}\mathbf{y}, \mathbf{V}_p).
\end{aligned}$$

■

Proof of Lemma 10

- Noting that

$$\begin{aligned} & l_t((\sigma\mathbf{B}-\mathbf{A})^{-1}\mathbf{B}\mathbf{y}_t) \\ = \; & \frac{1}{2}\mathbf{y}_t^\top\mathbf{B}(\sigma\mathbf{B}-\mathbf{A})^{-1}(\sigma\mathbf{B}-\mathbf{A})(\sigma\mathbf{B}-\mathbf{A})^{-1}\mathbf{B}\mathbf{y}_t - \mathbf{y}_t^\top\mathbf{B}(\sigma\mathbf{B}-\mathbf{A})^{-1}\mathbf{B}\mathbf{y}_t \\ = \; & -\frac{1}{2}\mathbf{y}_t^\top\mathbf{B}(\sigma\mathbf{B}-\mathbf{A})^{-1}\mathbf{B}\mathbf{y}_t, \end{aligned}$$

we have that

$$\begin{aligned} & \frac{1}{2}\|\mathbf{y}-(\sigma\mathbf{B}-\mathbf{A})^{-1}\mathbf{B}\mathbf{y}_t\|_{\mathbf{C}}^2 \\ = \; & \frac{1}{2}(\mathbf{y}-(\sigma\mathbf{B}-\mathbf{A})^{-1}\mathbf{B}\mathbf{y}_t)^\top(\sigma\mathbf{B}-\mathbf{A})(\mathbf{y}-(\sigma\mathbf{B}-\mathbf{A})^{-1}\mathbf{B}\mathbf{y}_t) \\ = \; & \frac{1}{2}\mathbf{y}^\top(\sigma\mathbf{B}-\mathbf{A})\mathbf{y} - \mathbf{y}_t^\top\mathbf{B}\mathbf{y} + \frac{1}{2}\mathbf{y}_t^\top\mathbf{B}(\sigma\mathbf{B}-\mathbf{A})^{-1}\mathbf{B}\mathbf{y}_t \\ = \; & l_t(\mathbf{y}) - l_t((\sigma\mathbf{B}-\mathbf{A})^{-1}\mathbf{B}\mathbf{y}_t) = \epsilon_t(\mathbf{y}). \end{aligned}$$

Let $r(\gamma)=\epsilon_t(\gamma\mathbf{y}_t)$. Since $\mathbf{y}_t\in \mathrm{St}_{\sigma\mathbf{B}-\mathbf{A}}(n,1)$, it holds that

$$\begin{aligned} r(\gamma) &= l_t(\gamma\mathbf{y}_t) - l_t((\sigma\mathbf{B}-\mathbf{A})^{-1}\mathbf{B}\mathbf{y}_t) \\ &= \frac{\gamma^2}{2}\mathbf{y}_t^\top(\sigma\mathbf{B}-\mathbf{A})\mathbf{y}_t - \gamma\mathbf{y}_t^\top\mathbf{B}\mathbf{y}_t - l_t((\sigma\mathbf{B}-\mathbf{A})^{-1}\mathbf{B}\mathbf{y}_t) \\ &= \frac{\gamma^2}{2} - \gamma\|\mathbf{y}_t\|_{\mathbf{B}}^2 - l_t((\sigma\mathbf{B}-\mathbf{A})^{-1}\mathbf{B}\mathbf{y}_t). \end{aligned}$$

$r(\gamma)$ is minimized at the root of $r'(\gamma)=0$, i.e., $\gamma=\|\mathbf{y}_t\|_{\mathbf{B}}^2$, and thus we have

$$\begin{aligned} \epsilon_t(\|\mathbf{y}_t\|_{\mathbf{B}}^2\mathbf{y}_t) &\le l_t(\mu_1\mathbf{y}_t) - l_t((\sigma\mathbf{B}-\mathbf{A})^{-1}\mathbf{B}\mathbf{y}_t) \\ &= \frac{\mu_1^2}{2}\mathbf{y}_t^\top(\sigma\mathbf{B}-\mathbf{A})\mathbf{y}_t - \mu_1\mathbf{y}_t^\top\mathbf{B}\mathbf{y}_t + \frac{1}{2}\mathbf{y}_t^\top\mathbf{B}(\sigma\mathbf{B}-\mathbf{A})^{-1}\mathbf{B}\mathbf{y}_t \\ &= \frac{\mu_1^2}{2}\mathbf{y}_t^\top\mathbf{B}^{\frac{1}{2}}\mathbf{C}\mathbf{B}^{\frac{1}{2}}\mathbf{y}_t - \mu_1\mathbf{y}_t^\top\mathbf{B}\mathbf{y}_t + \frac{1}{2}\mathbf{y}_t^\top\mathbf{B}^{\frac{1}{2}}\mathbf{C}^{-1}\mathbf{B}^{\frac{1}{2}}\mathbf{y}_t. \end{aligned}$$

By Equations (31)-(32), it holds that

$$\mathbf{C}^{-1} = \mathbf{B}^{\frac{1}{2}}\sum_{i=1}^{n}\mu_i\mathbf{v}_i\mathbf{v}_i^\top\mathbf{B}^{\frac{1}{2}}.$$

Then

$$\begin{aligned}
\epsilon_t(\|\mathbf{y}_t\|_{\mathbf{B}}^2\mathbf{y}_t) &\leq \frac{\mu_1^2}{2}\sum_{i=1}^{n}\frac{(\mathbf{v}_i^\top\mathbf{B}\mathbf{y}_t)^2}{\mu_i} - \mu_1\sum_{i=1}^{n}(\mathbf{v}_i^\top\mathbf{B}\mathbf{y}_t)^2 + \frac{1}{2}\sum_{i=1}^{n}\mu_i(\mathbf{v}_i^\top\mathbf{B}\mathbf{y}_t)^2 \\
&= \sum_{i=1}^{n}(\frac{\mu_1^2}{2\mu_i} - \mu_1 + \frac{1}{2}\mu_i)(\mathbf{v}_i^\top\mathbf{B}\mathbf{y}_t)^2 = \sum_{i=1}^{n}\frac{(\mu_1-\mu_i)^2}{2\mu_i}(\mathbf{v}_i^\top\mathbf{B}\mathbf{y}_t)^2 \\
&= \sum_{i=1}^{n}\frac{(\mu_1-\mu_i)^2}{2\mu_i}(\mathbf{v}_i^\top\mathbf{B}^{1/2}\mathbf{C}^{-1/2}\mathbf{C}^{1/2}\mathbf{B}^{1/2}\mathbf{y}_t)^2 \\
&= \sum_{i=1}^{n}\frac{(\mu_1-\mu_i)^2}{2\mu_i}(\sqrt{\mu_i}\mathbf{v}_i^\top\mathbf{B}^{1/2}\mathbf{C}^{1/2}\mathbf{B}^{1/2}\mathbf{y}_t)^2
\end{aligned}$$

$$\begin{aligned}
&= \frac{1}{2}\sum_{i=1}^{n}(\mu_1-\mu_i)^2(\mathbf{v}_i^\top\mathbf{B}^{1/2}\mathbf{C}^{1/2}\mathbf{B}^{1/2}\mathbf{y}_t)^2 = \frac{1}{2}\sum_{i>p}^{n}(\mu_1-\mu_i)^2(\mathbf{v}_i^\top\mathbf{B}^{1/2}\mathbf{C}^{1/2}\mathbf{B}^{1/2}\mathbf{y}_t)^2 \\
&\leq \frac{\mu_1^2}{2}\sum_{i=p+1}^{n}(\mathbf{v}_i^\top\mathbf{B}^{1/2}\mathbf{C}^{1/2}\mathbf{B}^{1/2}\mathbf{y}_t)^2 = \frac{\mu_1^2}{2}(1-\sum_{i=1}^{p}(\mathbf{v}_i^\top\mathbf{B}\mathbf{B}^{-1/2}\mathbf{C}^{1/2}\mathbf{B}^{1/2}\mathbf{y}_t)^2) \\
&= \frac{\mu_1^2}{2}\sin^2\theta(\mathbf{B}^{-1/2}\mathbf{C}^{1/2}\mathbf{B}^{1/2}\mathbf{y}_t, \mathbf{V}_p).
\end{aligned}$$

- It suffices for us to bound the extreme eigenvalues of $\sigma\mathbf{B}-\mathbf{A}$ for the complexity, as the Hessian of $l_t(\mathbf{y})$ satisfies

$$\lambda_{\min}(\sigma\mathbf{B}-\mathbf{A})\mathbf{I} \preccurlyeq \text{Hessian}(l_t(\mathbf{y})) = \sigma\mathbf{B}-\mathbf{A} \preccurlyeq \lambda_{\max}(\sigma\mathbf{B}-\mathbf{A})\mathbf{I}.$$

We can write that

$$\begin{aligned}
\lambda(\sigma\mathbf{B}-\mathbf{A}) &= \frac{1}{\lambda((\sigma\mathbf{B}-\mathbf{A})^{-1})} = \frac{1}{\lambda(\mathbf{B}^{-\frac{1}{2}}\mathbf{C}^{-1}\mathbf{B}^{-\frac{1}{2}})} \\
&\leq \frac{1}{\lambda_{\min}(\mathbf{B}^{-\frac{1}{2}})\lambda_{\min}(\mathbf{C}^{-1})\lambda_{\min}(\mathbf{B}^{-\frac{1}{2}})} \\
&= \frac{\lambda_{\max}(\mathbf{B})}{\lambda_{\min}(\mathbf{C}^{-1})} = \frac{\lambda_{\max}(\mathbf{B})}{\mu_d}
\end{aligned}$$

and

$$\begin{aligned}
\lambda(\sigma\mathbf{B}-\mathbf{A}) &= \frac{1}{(\sigma\mathbf{B}-\mathbf{A})^{-1})} = \frac{1}{\lambda(\mathbf{B}^{-\frac{1}{2}}\mathbf{C}^{-1}\mathbf{B}^{-\frac{1}{2}})} \\
&\geq \frac{1}{\lambda_{\max}(\mathbf{B}^{-\frac{1}{2}})\lambda_{\max}(\mathbf{C}^{-1})\lambda_{\max}(\mathbf{B}^{-\frac{1}{2}})} \\
&= \frac{\lambda_{\min}(\mathbf{B})}{\lambda_{\max}(\mathbf{C}^{-1})} = \frac{\lambda_{\min}(\mathbf{B})}{\mu_1}.
\end{aligned}$$

Thus, the ratio of the smooth parameter to the strong convexity parameter for $l_t(\mathbf{y})$ can be bounded by

$$\frac{\frac{\lambda_{\max}(\mathbf{B})}{\mu_d}}{\frac{\lambda_{\min}(\mathbf{B})}{\mu_1}} = \kappa(\mathbf{B})\frac{\lambda_1}{\Delta_p}.$$

■