Faster Noisy Power Method

Zhiqiang Xu XUZHIQIANG04@BAIDU.COM
Cognitive Computing Lab
Baidu Research

Ping Li LIPING11@BAIDU.COM
Cognitive Computing Lab
Baidu Research

10900 NE 8th St. Bellevue, WA 98004, USA

Editors: Sanjoy Dasgupta and Nika Haghtalab

Abstract

Given the capability to handle diverse resource constraints, such as communication, memory, or privacy, the noisy power method (Hardt and Price, 2014), as a meta algorithm for computing the dominant eigenspace of a matrix, has found wide applications in data analysis and statistics (e.g., PCA). For an input data matrix, the performance of the algorithm, as with the noiseless case, is characterized by the spectral gap, which largely dictates the convergence rate and affects the noise tolerance level as well (Hardt and Price, 2014). A recent analysis (Balcan et al., 2016) improved the dependency over the consecutive spectral gap $(\lambda_k - \lambda_{k+1})$ (Hardt and Price, 2014) to the dependency over $(\lambda_k - \lambda_{q+1})$, where q could be much greater than the target rank k and thus result in better performance by a significantly larger gap. However, $(\lambda_k - \lambda_{q+1})$ could still be quite small and potentially limit the applicability. In this paper, we further improve the dependency of the convergence rate over $O(\lambda_k - \lambda_{q+1})$ to dependency over $\tilde{O}(\sqrt{\lambda_k - \lambda_{q+1}})$ in a certain regime of a new parameter, for a faster noise-tolerant algorithm. To achieve this goal, we propose faster noisy power method which introduces the momentum acceleration into the noisy power iteration, and present a novel analysis that differs from previous ones (Hardt and Price, 2014; Balcan et al., 2016). We also extend our algorithm to the distributed PCA and memory-efficient streaming PCA and get improved results accordingly in terms of the gap dependence.

Keywords: Noisy power method, spectral gap dependency, momentum acceleration

1. Introduction

In data analysis and statistics, it often needs to find the dominant eigenspace of a matrix, which can be done by the classic power method or Krylov subspace method (e.g., Lanczos algorithm, known as a faster counterpart of the power method) (Golub and Van Loan, 2013). One of emerging trends is that many applications arising recently require to approximately compute dominant eigenspace in the presence of noise of various forms such as missing entries, sampling error, approximation error, privacy constraint, or adversarial attack (Mitliagkas et al., 2013; Hardt and Roth, 2013; Hardt and Price, 2014; Liu et al., 2015; Xu and Li, 2019, 2020, 2021). In this case, the noisy power method (Hardt and Price, 2014) turns out to be a meta algorithm that can fulfill a wide range of resource constraints like the above, given that it is a fast general purpose method for the dominant eigenspace computation under noise-corrupted matrix-vector multiplications. As with the noiseless setting, the convergence rate of the noisy power method is inversely proportional to and largely

dominated by the spectral gap. Particularly, Hardt and Price (2014) showed the dependence of the convergence on the consecutive spectral gap $(\lambda_k - \lambda_{k+1})$, where λ_i represents the i-th largest eigenvalue of the given real symmetric data matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ and k is the target rank (i.e., the goal to find a top-k eigenspace of $\mathbf{A}$, denoted as $\mathbf{U}_k$). However, the consecutive spectral gaps are too small for real large data (Musco and Musco, 2015). For faster convergence and numerical stability, in practice, one can use an iteration rank p (i.e., matrix iterate $\mathbf{X}_t \in \mathbb{R}^{n\times p}$) larger than the target rank k (Hardt and Price, 2014; Musco and Musco, 2015; Wang et al., 2015). This implies the possibility of dependence on a larger spectral gap $(\lambda_k - \lambda_{p+1})$, which was conjectured in Hardt and Price (2014). This remedy can't be theoretically justified by their proof technique even in the noiseless setting until Gu (2015) and Balcan et al. (2016). In the noiseless setting, Gu (2015) presented a theoretical justification for the remedy of a larger iteration rank with a rigourous proof that under mild conditions the dependency over $(\lambda_k - \lambda_{k+1})$ can be improved to be over $(\lambda_k - \lambda_{q+1})$ for some $k \le q \le p$, which may be significantly greater. Given the increasing practical value of the noise setting and growing attention from both machine learning and theoretical computer science communities, Balcan et al. (2016) further showed similar results for the noisy power method by a different analysis.

In fact, the spectral gap $(\lambda_k - \lambda_{q+1})$, albeit larger than the consecutive one, could still be small and limit the applicability of the noisy power method. One natural question then is:

Can we further improve the dependency over the spectral gap $(\lambda_k - \lambda_{q+1})$*?*

Indeed, Balcan et al. (2016) noted that Krylov iteration (Golub and Van Loan, 2013) has an improved dependency in the noiseless setting (Musco and Musco, 2015), but only expected interesting results for the noisy Krylov subspace method without further analysis. This noisy Krylov iteration seems quite difficult to analyze directly. In this work, we equip the noisy power method with momentum acceleration (Xu et al., 2018) instead to give a faster noisy power method:

$$\mathbf{X}_{t+1}\mathbf{R}_{t+1} = \mathbf{A}\mathbf{X}_t - \beta\mathbf{X}_{t-1}\mathbf{R}_t^{-1} + \boldsymbol{\xi}_t \in \mathbb{R}^{n\times p}, \tag{1}$$

where $\beta > 0$ is the momentum parameter, $\boldsymbol{\xi}_t$ is the noise matrix for iteration t, and the left-hand side represents the QR factorization (Golub and Van Loan, 2013) of the right-hand side such that $\mathbf{X}_t \in \mathbb{R}^{n\times p}$ remains column-orthonormal and $\mathbf{R}_t \in \mathbb{R}^{p\times p}$. The initials include $\mathbf{X}_{-1} = \mathbf{0} \in \mathbb{R}^{n\times p}$, random column-orthonormal $\mathbf{X}_0 \in \mathbb{R}^{n\times p}$ obtained from the QR factorization of an entrywise i.i.d. standard normal matrix $\widehat{\mathbf{X}}_0 \in \mathbb{R}^{n\times p}$, i.e., $\mathbf{X}_0\mathbf{R}_0 = \widehat{\mathbf{X}}_0$. When $\beta = 0$, Eq.(1) recovers the noisy power method (Hardt and Price, 2014; Balcan et al., 2016). Note that the noise matrix $\boldsymbol{\xi}_t$ could model a variety of resource constraints as mentioned previously, including stochastic sampling errors considered in Xu et al. (2018). The analysis in Xu et al. (2018) is specific to $\boldsymbol{\xi}_t \equiv \mathbf{0}$ with $p = k$ or stochastic errors with $p = k = 1$, and is inapplicable to our general noise model with a larger iteration rank $p > k \ge 1$ than the target rank k.

We give a novel analysis different from Hardt and Price (2014); Gu (2015); Balcan et al. (2016) to show an improved spectral gap dependency of the convergence of our faster noisy power method in Eq. (1) over the noisy power method, i.e., $\tilde{O}(\sqrt{\lambda_k - \lambda_{q+1}})$. To analyze the convergence of Eq. (1), we first have it paired with the identity equation $\mathbf{X}_t = \mathbf{X}_t$ to get an equivalent augmented update equation of iterate $\mathbf{Y}_t \in \mathbb{R}^{2n\times p}$ formed by two consecutive original iterates $\mathbf{X}_t$ and $\mathbf{X}_{t-1}$, in a form very similar to the noisy power method. Despite a similar form, a key difference from the previous noisy power method (Hardt and Price, 2014; Gu, 2015; Balcan et al., 2016) lies at the asymmetry of the augmented data matrix $\mathbf{B} \in \mathbb{R}^{2n\times 2n}$ associated with the matrix-vector product, which causes

significant complications of analysis due to the Schur decomposition, instead of the commonly used eigenvalue decomposition for real symmetric matrices, as we shall see in the sequel. We then need to show the rate of the convergence of $\mathbf{Y}_t$, or the newly minted and augmented noisy power iteration, to the dominant k-dimensional invariant subspace of $\mathbf{B}$ (denoted as $\mathbf{V}_k$). To this end, we build upon the so-called rank-k perturbation on $\mathbf{V}_q$ by $\mathbf{Y}_t$, i.e., h_t, proposed in Balcan et al. (2016), which is a characterization between a rank-p subspace $\mathbf{Y}_t$ and the rank-k target subspace $\mathbf{V}_k$ through an intermediate subspace $\mathbf{V}_q$. We adapt the rank-k perturbation on $\mathbf{V}_q$ by $\mathbf{Y}_t$ to our case and denote it as our potential function Φ_t. Except for that, however, the proof idea of Balcan et al. (2016) can't be applied in our case, as it will lead to an even worse noise tolerance bound in terms of spectral gap dependence. In contrast to Balcan et al. (2016) showing the constant contraction of $h_t - O(\epsilon)$ over iterations, we directly establish Φ_t's geometric shrinkage across iterations under mild conditions on the momentum parameter, which turns out to give improved noise tolerance. Finally, based on the convergence of $\mathbf{Y}_t$, we show the rate of the convergence of $\mathbf{X}_t$ or Eq. (1) to $\mathbf{U}_k$.

We further apply our meta algorithm to the distributed PCA and memory-efficient streaming PCA, and get improved results accordingly in terms of the gap dependence.

2. Notions and Notations

Given a positive semi-definite data matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$, let $\mathbf{u}_j$ represent $\mathbf{A}$'s eigenvector of unit length corresponding to the j-th largest eigenvalue $\lambda_j \geq 0$. Denote $\mathbf{\Sigma}_j = \operatorname{diag}(\lambda_1, \cdots, \lambda_j), \mathbf{\Sigma}_{-j} = \operatorname{diag}(\lambda_{j+1}, \cdots, \lambda_n), \mathbf{U}_j = \big[\mathbf{u}_1, \cdots, \mathbf{u}_j\big]$, and $\mathbf{U}_{-j} = \big[\mathbf{u}_{j+1}, \cdots, \mathbf{u}_n\big]$. For any $j \in \{1, 2, \cdots, n\}$, $\mathbf{A} = \big[\mathbf{U}_j, \mathbf{U}_{-j}\big] \operatorname{diag}(\mathbf{\Sigma}_j, \mathbf{\Sigma}_{-j}) \big[\mathbf{U}_j, \mathbf{U}_{-j}\big]^\top$ is $\mathbf{A}$'s full eigenvalue decomposition, where $\big[\mathbf{U}_j, \mathbf{U}_{-j}\big]$ is orthogonal. With a bit abuse of notation, we also use $\mathbf{U}_j$ to represent its column space for brevity, e.g., $\mathbf{U}_k$ is a top-k eigenspace of $\mathbf{A}$. For later convenience, let i exclusively represent the imaginary unit equal to $\sqrt{-1}$, $\overline{\square}$ solely represent the conjugate of a matrix, and $\square^{\mathrm{H}} \triangleq \overline{\square}^\top$ the conjugate transpose of a matrix, throughout the paper. $\mathbf{I}_j$ represents the identity matrix of size $j \times j$, and the identity matrix of appropriate size if the subscript is missing (i.e., $\mathbf{I}$). In addition, $\square^\dagger$ represents the pseudo-inverse of a matrix, and $\operatorname{col}(\square)$ stands for the column space of a matrix. Let $\mathbf{Z}_1 \in \mathbb{C}^{m\times d_1}, \mathbf{Z}_2 \in \mathbb{C}^{m\times d_2}$ ($d_j \leq m$ for $j = 1, 2$) and $d = \min\{d_1, d_2\}$. The vector of principal angles between $\mathbf{Z}_1$ and $\mathbf{Z}_2$ then is defined as $\boldsymbol{\theta}(\mathbf{Z}_1, \mathbf{Z}_2) \triangleq \big[\cos^{-1}\big(\sigma_d(\mathbf{Z}_1^{\mathrm{H}}\mathbf{Z}_2)\big), \cdots, \cos^{-1}\big(\sigma_1(\mathbf{Z}_1^{\mathrm{H}}\mathbf{Z}_2)\big)\big]^\top$, where $\sigma_j(\square)$ represents the j-th largest singular value of a matrix, and naturally $\frac{\pi}{2} \geq \theta_1(\mathbf{Z}_1, \mathbf{Z}_2) \geq \cdots \geq \theta_d(\mathbf{Z}_1, \mathbf{Z}_2) \geq 0$. In addition, let $\theta_{\max}(\mathbf{Z}_1, \mathbf{Z}_2) \triangleq \theta_1(\mathbf{Z}_1, \mathbf{Z}_2)$, and $\theta_{\min}(\mathbf{Z}_1, \mathbf{Z}_2) \triangleq \theta_d(\mathbf{Z}_1, \mathbf{Z}_2)$. Distance between subspaces can be characterized by their largest principal angle, i.e., $\theta_{\max}(\cdot, \cdot)$.

3. Faster Noisy Power Method

The pseudo code of the faster noisy power method is described in Algorithm 1. In this section, we give our main results and proof for Algorithm 1. The missing proofs can be found in Appendix.

3.1. Main Results

Theorem 1 *Let $k \leq q \leq p$ and assume that $\lambda_q > 2\sqrt{\beta} \geq \lambda_{q+1}$ for $\mathbf{A} \in \mathbb{R}^{n\times n} \succcurlyeq \mathbf{0}$. Define*

$$\mathbf{B} = \begin{bmatrix} \mathbf{A} & -\beta\mathbf{I} \\ \mathbf{I} & \mathbf{0} \end{bmatrix}, \quad \mathbf{Y}_t = \begin{bmatrix} \mathbf{X}_t \\ \mathbf{X}_{t-1}\mathbf{R}_t^{-1} \end{bmatrix} \mathbf{S}_t, \quad \mathbf{E}_t = \begin{bmatrix} \boldsymbol{\xi}_t \\ \mathbf{0} \end{bmatrix} \mathbf{S}_t, \quad \mathbf{S}_t = \begin{cases} (\mathbf{I} + \mathbf{R}_t^{-\top}\mathbf{R}_t^{-1})^{-\frac{1}{2}}, & t \geq 1 \\ \mathbf{I}, & t = 0 \end{cases}.$$

Algorithm 1 FNPM

1: **Input:** positive semi-definite matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$, momentum parameter $\beta > 0$, target rank k, iteration rank $p \geq k$, iteration number T.
2: **Output:** approximate top-k eigenspace $\mathbf{X}_T$.
3: Set $\mathbf{X}_{-1} = \mathbf{0} \in \mathbb{R}^{n\times p}$ and QR factorize an entry-wise i.i.d. standard Gaussian matrix $\widehat{\mathbf{X}}_0 \in \mathbb{R}^{n\times p}$ such that $\mathbf{X}_0\mathbf{R}_0 = \widehat{\mathbf{X}}_0$
4: **for** $t = 0, 1, \cdots, T-1$ **do**
5: $\quad \widehat{\mathbf{X}}_{t+1} = \mathbf{A}\mathbf{X}_t - \beta\mathbf{X}_{t-1}\mathbf{R}_t^{-1} + \boldsymbol{\xi}_t$ for some noise matrix $\boldsymbol{\xi}_t$
6: $\quad$ QR factorize $\widehat{\mathbf{X}}_{t+1}$ such that $\mathbf{X}_{t+1}\mathbf{R}_{t+1} = \widehat{\mathbf{X}}_{t+1}$
7: **end for**

If the noise matrix $\boldsymbol{\xi}_t \in \mathbb{R}^{n\times p}$ *satisfies that* $\|\boldsymbol{\xi}_t\|_2 \leq \frac{1}{256}\frac{(\lambda_k^+ - \sqrt{\beta})^2}{(1+l_1)(1+\beta)}\sin\theta_{\max}(\mathbf{Y}_t, \mathbf{V}_q)$,

$$\|\mathbf{U}_q^\top \boldsymbol{\xi}_t\|_2 \leq \frac{1}{16}\frac{\lambda_q^+(\lambda_k^+ - \sqrt{\beta})}{(1+l_2)(1+\lambda_k^+)}\cos\theta_{\max}(\mathbf{Y}_t, \mathbf{W}_q),$$

and

$$\mathrm{col}((\mathbf{W}_q^{\mathrm{H}}\mathbf{Y}_t + \mathrm{diag}^{-1}(\lambda_1^+, \cdots, \lambda_q^+)\mathbf{W}_q^{\mathrm{H}}\mathbf{E}_t)^\dagger \begin{bmatrix}\mathbf{I}_k \\ \mathbf{0}\end{bmatrix}) \subset \mathrm{col}((\mathbf{W}_q^{\mathrm{H}}\mathbf{Y}_t)^\dagger \begin{bmatrix}\mathbf{I}_k \\ \mathbf{0}\end{bmatrix}), \tag{2}$$

where $\mathbf{V}_q$ *and* $\mathbf{W}_q$ *are the dominant* q*-dimensional invariant subspaces of* $\mathbf{B}$ *and* $\mathbf{B}^\top$*, respectively, whose eigenvalues are* $\lambda_s^\pm = \frac{\lambda_s \pm \sqrt{\lambda_s^2 - 4\beta}}{2}$ *(see Lemma 2), and* $l_1 = \max_{0\leq t\leq T}\frac{\sin\theta_{\max}(\mathbf{Y}_t,\mathbf{V}_q)}{\sin\theta_{\min}(\mathbf{Y}_t,\mathbf{V}_q)}$, $l_2 = \max_{0\leq t\leq T}\frac{\cos\theta_{\min}(\mathbf{Y}_t,\mathbf{W}_q)}{\cos\theta_{\max}(\mathbf{Y}_t,\mathbf{W}_q)}$*, then after Algorithm 1 runs for*

$$T \geq 4\sqrt{\frac{\lambda_k}{\lambda_k - 2\sqrt{\beta}}}\log\Big(\sqrt{\frac{1+(\lambda_k^+)^2}{(\lambda_k^+)^2}\frac{\beta^2+(\lambda_k^+)^2}{(\lambda_k^+)^2}}\frac{1+\beta+\lambda_q}{\sqrt{\lambda_q^2-4\beta}}\frac{1+\beta}{\lambda_k^+ - \sqrt{\beta}}\frac{32k\tau\sqrt{n}}{\sqrt{p}-\sqrt{q-1}}\frac{1}{\epsilon}\Big)$$

iterations, we have that $\sin\theta_{\max}(\mathbf{X}_T, \mathbf{U}_k) < \epsilon$ *with probability at least* $1 - \tau^{-\Omega(p+1-q)} - e^{-\Omega(n)}$.

Remark 1 Note in the above theorem that $\lambda_k^+ - \sqrt{\beta} = O(\sqrt{\lambda_k - \lambda_{q+1}})$ when $2\sqrt{\beta}$ is close to λ_{q+1} (see the proof of Theorem 1). Table 1 shows the comparison of our results with existing ones (Hardt and Price, 2014; Balcan et al., 2016). When $2\sqrt{\beta}$ is close to λ_{q+1}, our convergence rate $T = \tilde{O}(\sqrt{\frac{\lambda_k}{\lambda_k - \lambda_{q+1}}})$, up to log factors, which significantly improves over $O(\frac{\lambda_k}{\lambda_k - \lambda_{q+1}}\log\frac{1}{\epsilon})$ (Balcan et al., 2016). Our noise tolerance also has significant improvement on the spectral gap dependency, i.e., $\tilde{O}(\sqrt{\lambda_k - \lambda_{q+1}})$, over $\tilde{O}(\lambda_k - \lambda_{q+1})$ (Balcan et al., 2016), in terms of $\|\mathbf{U}_q^\top\boldsymbol{\xi}_t\|_2$. But in terms of $\|\boldsymbol{\xi}_t\|_2$, it has the same dependency as Balcan et al. (2016), i.e., $O(\lambda_k - \lambda_{q+1})$. Due to $\|\mathbf{U}_q^\top\boldsymbol{\xi}_t\|_2 \leq \|\boldsymbol{\xi}_t\|_2$, this may worsen the gap dependence of $\|\mathbf{U}_q^\top\boldsymbol{\xi}_t\|_2$ to be the same as that of $\|\boldsymbol{\xi}_t\|_2$ when $\sqrt{\lambda_k - \lambda_{q+1}}$ is very small. Nevertheless, in general, $\sin\theta_{\max}(\mathbf{Y}_t, \mathbf{V}_q)$ and $\cos\theta_{\max}(\mathbf{Y}_t, \mathbf{W}_q)$ are significantly larger than ϵ before convergence. Thus, overall our noise tolerance bounds are better than those of Balcan et al. (2016). In addition, another noise condition in Eq. (2), which is important to achieve geometric shrinkage across iterations without needing the noise level to be as small as the accuracy, seems quite restrictive, but it should be satisfied by two settings considered in Section 4, i.e., distributed or streaming PCA, because $\boldsymbol{\xi}_t = \mathbf{0}$ or $\boldsymbol{\xi}_t = (\mathbf{A}_t - \mathbf{A})\mathbf{X}_t$.

Table 1: Comparison with existing results.

	T	$\|\boldsymbol{\xi}_t\|_2$	$\|\mathbf{U}_k^\top \boldsymbol{\xi}_t\|_2$ or $\|\mathbf{U}_q^\top \boldsymbol{\xi}_t\|_2$
Hardt and Price (2014)	$\tilde{O}(\frac{\lambda_k}{\lambda_k-\lambda_{k+1}})$	$O((\lambda_k-\lambda_{k+1})\epsilon)$	$O((\lambda_k-\lambda_{k+1})\cos\theta_{\max}(\mathbf{X}_0,\mathbf{U}_k))$
Balcan et al. (2016)	$\tilde{O}(\frac{\lambda_k}{\lambda_k-\lambda_{q+1}})$	$O((\lambda_k-\lambda_{q+1})\epsilon)$	$O((\lambda_k-\lambda_{q+1})\epsilon\cos\theta_{\max}(\mathbf{X}_0,\mathbf{U}_q))$
Ours (when $2\sqrt{\beta}$ is close to λ_{q+1})	$\tilde{O}(\sqrt{\frac{\lambda_k}{\lambda_k-\lambda_{q+1}}})$	$O((\lambda_k-\lambda_{q+1})\sin\theta_{\max}(\mathbf{Y}_t,\mathbf{V}_q))$	$O(\sqrt{\lambda_k-\lambda_{q+1}}\cos\theta_{\max}(\mathbf{Y}_t,\mathbf{W}_q))$

Remark 2 To further understand the improvement of the convergence rate, we revisit the power law example with decaying spectrum given in Balcan et al. (2016), i.e., $\lambda_k \asymp k^{-\alpha}$ for some parameter $\alpha > 1$. This spectral decay property is common in many data matrices arising in practical data applications (Liu et al., 2015). By setting $q = ck$ for some constant $c > 1$, the linear dependence on the relative spectral gap gives factor $\frac{\lambda_k-\lambda_{q+1}}{\lambda_k} = 1 - c^{-\alpha}$ for the noisy power method, and thus total number of flops is $\tilde{O}(\frac{nk^2}{1-c^{-\alpha}})$. For our FNPM, the nearly square-root dependence gives factor close to $\sqrt{1-c^{-\alpha}}$ which could be much larger than $(1-c^{-\alpha})$ and then yield a much smaller total number of flops at $\tilde{O}(\frac{nk^2}{\sqrt{1-c^{-\alpha}}})$.

Remark 3 The convergence results for other values of β are easy to get by the proof of Theorem 1 with a slight change (see Remark 5). There are three cases: 1) when $2\sqrt{\beta} \geq \lambda_k$, Algorithm 1 may not converge; 2) when $\lambda_k > 2\sqrt{\beta} \geq \lambda_q$, $\sin\theta_{\max}(\mathbf{X}_T,\mathbf{U}_k) < \epsilon$ holds for $T = \tilde{O}(\sqrt{\frac{\lambda_k}{\lambda_k-2\sqrt{\beta}}})$ (the same form as Theorem 1); 3) when $\lambda_{q+1} > 2\sqrt{\beta}$, $\sin\theta_{\max}(\mathbf{X}_T,\mathbf{U}_k) < \epsilon$ holds for $T = \tilde{O}(\frac{\lambda_k^+}{\lambda_k^+-\lambda_{q+1}^+})$. Note that we always have $\frac{\lambda_k^+}{\lambda_k^+-\lambda_{q+1}^+} < \frac{\lambda_k}{\lambda_k-\lambda_{q+1}}$. The noise tolerance can be derived accordingly as well.

Remark 4 There is no need to worry about if we will get $2\sqrt{\beta} \geq \lambda_q$ which makes Algorithm 1 either possibly diverge or converge slowly (how slow it is depends on how close $2\sqrt{\beta}$ is to λ_k). This is because we could simply use varying β by setting $2\sqrt{\beta_t}$ to be the $(q+1)$-th largest diagonal entry of $\widehat{\boldsymbol{\Sigma}}_p = \mathbf{X}_t^\top(\mathbf{A}\mathbf{X}_t + \boldsymbol{\xi}_t) \in \mathbb{R}^{p\times p}$ and $2\sqrt{\beta_t} \leq \lambda_{q+1}$ always holds approximately. However, it is difficult to set $\boldsymbol{\xi}$ to meet the noise condition in practice, because it involves not only a lot of ground truth information but also the current progress of the iterate to the solution. Nonetheless, our results show that the noise could be much larger at initial stage than later stage, but eventually the noise needs to be at least as small as the desired accuracy in order for convergence. Moreover, in a safe noise range, our algorithm could converge faster than the plain noisy power method (i.e., $\beta = 0$).

3.2. Proof of Theorem 1

We start from pairing Eq. (1) with the identity equation $\mathbf{X}_t = \mathbf{X}_t$ to get an augmented update equation:

$$\begin{bmatrix} \mathbf{X}_{t+1} \\ \mathbf{X}_t\mathbf{R}_{t+1}^{-1} \end{bmatrix} \mathbf{R}_{t+1} = \begin{bmatrix} \mathbf{A} & -\beta\mathbf{I} \\ \mathbf{I} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{X}_t \\ \mathbf{X}_{t-1}\mathbf{R}_t^{-1} \end{bmatrix} + \begin{bmatrix} \boldsymbol{\xi}_t \\ \mathbf{0} \end{bmatrix}.$$

Post-multiplying both sides of the equation above by $\mathbf{S}_t$ and letting $\widehat{\mathbf{R}}_t = \mathbf{S}_t^{-1}\mathbf{R}_t\mathbf{S}_{t-1}$ for $t \geq 1$, with notations given in Theorem 1, we arrive at an equivalent update equation of the same form as the noisy power iteration:

$$\mathbf{Y}_{t+1}\widehat{\mathbf{R}}_{t+1} = \mathbf{B}\mathbf{Y}_t + \mathbf{E}_t, \tag{3}$$

where $\mathbf{Y}_t$ is column-orthonormal, i.e., $\mathbf{Y}_t^\top \mathbf{Y}_t = \mathbf{I}$. Despite the simple form, however, the established theories for the noisy power method in Hardt and Price (2014); Balcan et al. (2016) can't be applied here, as they only work for real symmetric target matrices (e.g., $\mathbf{A}$). In our case, the target matrix $\mathbf{B}$ is asymmetric. To continue, we need the following lemma about the Schur decompositions of real asymmetric matrices $\mathbf{B}$ and $\mathbf{B}^\top$:

Lemma 2 *If there exists an integer j such that $\lambda_j > 2\sqrt{\beta} \geq \lambda_{j+1}$, then we have the following Schur decompositions:*

$$\mathbf{B} = \mathbf{J}(1, -1, -1), \quad \mathbf{B}^\top = \mathbf{J}(\beta, \beta, 1),$$

where

$$\mathbf{J}(a,b,c) = \begin{bmatrix} \mathbf{U}_j\mathbf{D}_j(a) & \mathbf{U}_{-j}\mathbf{D}_{-j}(a) & c\mathbf{U}_{-j}\mathbf{K}_{-j}(a) & c\mathbf{U}_j\mathbf{K}_j(a) \\ -c\mathbf{U}_j\mathbf{K}_j(a) & -c\mathbf{U}_{-j}\mathbf{K}_{-j}(a) & \mathbf{U}_{-j}\overline{\mathbf{D}_{-j}(a)} & \mathbf{U}_j\overline{\mathbf{D}_j(a)} \end{bmatrix} \times$$
$$\begin{bmatrix} \mathbf{\Sigma}_j^+ & \mathbf{0} & \mathbf{0} & \mathbf{F}(c) \\ \mathbf{0} & \mathbf{\Sigma}_{-j}^+ & \mathbf{T}(b) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{\Sigma}_{-j}^- & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{\Sigma}_j^- \end{bmatrix} \begin{bmatrix} \mathbf{U}_j\mathbf{D}_j(a) & \mathbf{U}_{-j}\mathbf{D}_{-j}(a) & c\mathbf{U}_{-j}\mathbf{K}_{-j}(a) & c\mathbf{U}_j\mathbf{K}_j(a) \\ -c\mathbf{U}_j\mathbf{K}_j(a) & -c\mathbf{U}_{-j}\mathbf{K}_{-j}(a) & \mathbf{U}_{-j}\overline{\mathbf{D}_{-j}(a)} & \mathbf{U}_j\overline{\mathbf{D}_j(a)} \end{bmatrix}^{\mathrm{H}}$$

with notations

$$\lambda_s^\pm = \frac{\lambda_s \pm \sqrt{\lambda_s^2 - 4\beta}}{2}, \qquad \begin{aligned} \mathbf{\Sigma}_j^+ &= \mathrm{diag}(\lambda_1^+, \cdots, \lambda_j^+), & \mathbf{\Sigma}_{-j}^+ &= \mathrm{diag}(\lambda_{j+1}^+, \cdots, \lambda_n^+), \\ \mathbf{\Sigma}_j^- &= \mathrm{diag}(\lambda_1^-, \cdots, \lambda_j^-), & \mathbf{\Sigma}_{-j}^- &= \mathrm{diag}(\lambda_{j+1}^-, \cdots, \lambda_n^-), \end{aligned}$$
$$\mathbf{D}_j(a) = \mathrm{diag}\Big(\frac{\lambda_1^+}{\sqrt{a^2+|\lambda_1^+|^2}}, \cdots, \frac{\lambda_j^+}{\sqrt{a^2+|\lambda_j^+|^2}}\Big), \quad \mathbf{D}_{-j}(a) = \mathrm{diag}\Big(\frac{\lambda_{j+1}^+}{\sqrt{a^2+|\lambda_{j+1}^+|^2}}, \cdots, \frac{\lambda_n^+}{\sqrt{a^2+|\lambda_n^+|^2}}\Big),$$
$$\mathbf{K}_j(a) = \mathrm{diag}\Big(\frac{a}{\sqrt{a^2+|\lambda_1^+|^2}}, \cdots, \frac{a}{\sqrt{a^2+|\lambda_j^+|^2}}\Big), \quad \mathbf{K}_{-j}(a) = \mathrm{diag}\Big(\frac{a}{\sqrt{a^2+|\lambda_{j+1}^+|^2}}, \cdots, \frac{a}{\sqrt{a^2+|\lambda_n^+|^2}}\Big),$$

$\mathbf{T}(b) = b\mathbf{I} + \frac{1}{b}(\mathbf{\Sigma}_{-j}^-)^2$, *and* $\mathbf{F}(c) = c(1+\beta)\mathbf{I}$. *If there exists any $s > j$ such that $\lambda_s = 2\sqrt{\beta}$, then $\lambda_s^\pm = \sqrt{\beta}$ and the corresponding diagonal entry in the block $\mathbf{T}(b)$ is replaced with $(1+\beta)$.*

Note that $\lambda_s^\pm$ is a conjugate pair of eigenvalues of $\mathbf{B}$ when $\lambda_s^2 < 4\beta$, and all the eigenvalues of $\mathbf{B}$ are on the diagonal of the upper triangular factor matrix of $\mathbf{J}(1,-1,-1)$, arranged in descending order of magnitude except for those with the lower right block $\mathbf{\Sigma}_j^-$ where ascending order applies. The above lemma is a unified form of Lemmas 2.1 and 2.2 given in Appendix about Schur decompositions of $\mathbf{B}$ and $\mathbf{B}^\top$, respectively. For other values of $\beta > 0$, i.e., $2\sqrt{\beta} \geq \lambda_1$ and $0 < 2\sqrt{\beta} < \lambda_n$, the Schur decompositions of $\mathbf{B}$ and $\mathbf{B}^\top$ are straightforward based on the proof of Lemma 2.

By Lemma 2 and the assumption that $\lambda_q > 2\sqrt{\beta} \geq \lambda_{q+1}$, the Schur decomposition of the real asymmetric matrix $\mathbf{B}$ can be rewritten as:

$$\mathbf{B} = \begin{bmatrix} \mathbf{V}_q & \mathbf{V}_{-q} \end{bmatrix} \begin{bmatrix} \mathbf{\Lambda}_q & \mathbf{\Lambda}_{q,-q} \\ \mathbf{0} & \mathbf{\Lambda}_{-q} \end{bmatrix} \begin{bmatrix} \mathbf{V}_q & \mathbf{V}_{-q} \end{bmatrix}^{\mathrm{H}}, \tag{4}$$

where $\mathbf{V}_{-q}$ represents the orthogonal complement of $\mathbf{V}_q$ in $\mathbb{C}^{2n}$, and $\mathbf{V}_q = \begin{bmatrix}\mathbf{U}_q\mathbf{D}_q(1)\\ \mathbf{U}_q\mathbf{K}_q(1)\end{bmatrix}$,

$$\mathbf{\Lambda}_q = \mathbf{\Sigma}_q^+, \ \ \mathbf{\Lambda}_{q,-q} = \begin{bmatrix}\mathbf{0} & \mathbf{0} & -(1+\beta)\mathbf{I}\end{bmatrix}, \ \ \mathbf{\Lambda}_{-q} = \begin{bmatrix}\mathbf{\Sigma}_{-q}^+ & \mathbf{T}(-1) & \mathbf{0}\\ \mathbf{0} & \mathbf{\Sigma}_{-q}^- & \mathbf{0}\\ \mathbf{0} & \mathbf{0} & \mathbf{\Sigma}_q^-\end{bmatrix}, \tag{5}$$

and $\mathbf{V}_q$ spans the unique dominant q-dimensional invariant subspace of $\mathbf{B}$ corresponding to top-q eigenvalues of $\mathbf{B}$ in magnitude which are on the diagonal of $\mathbf{\Sigma}_q^+$. Since $\mathbf{\Lambda}_q$ and $\mathbf{\Lambda}_{-q}$ don't have eigenvalues in common, by Lemma 7.1.5 in Golub and Van Loan (2013), there exists a matrix $\mathbf{\Omega} \in \mathbb{C}^{q\times(2n-q)}$ such that $\mathbf{\Lambda}_q\mathbf{\Omega} - \mathbf{\Omega}\mathbf{\Lambda}_{-q} = -\mathbf{\Lambda}_{q,-q}$ and thus

$$\begin{bmatrix}\mathbf{\Lambda}_q & \mathbf{\Lambda}_{q,-q}\\ \mathbf{0} & \mathbf{\Lambda}_{-q}\end{bmatrix} = \begin{bmatrix}\mathbf{I} & \mathbf{\Omega}\\ \mathbf{0} & \mathbf{I}\end{bmatrix}\begin{bmatrix}\mathbf{\Lambda}_q & \mathbf{0}\\ \mathbf{0} & \mathbf{\Lambda}_{-q}\end{bmatrix}\begin{bmatrix}\mathbf{I} & \mathbf{\Omega}\\ \mathbf{0} & \mathbf{I}\end{bmatrix}^{-1}. \tag{6}$$

Particularly, by Theorem 7.3.1 of Golub and Van Loan (2013), it holds that $\mathbf{B}^\top\widehat{\mathbf{W}}_q = \widehat{\mathbf{W}}_q\mathbf{\Lambda}_q$, where $\widehat{\mathbf{W}}_q = \mathbf{V}_q - \mathbf{V}_{-q}\mathbf{\Omega}^{\mathrm{H}}$. Thus, orthonormal $\mathbf{W}_q = \widehat{\mathbf{W}}_q(\mathrm{I} + \mathbf{\Omega}\mathbf{\Omega}^{\mathrm{H}})^{-\frac{1}{2}}$ spans the dominant q-dimensional invariant subspace of $\mathbf{B}^\top$. By Lemma 2, it then holds that

$$\mathrm{col}(\mathbf{W}_q) = \mathrm{col}(\begin{bmatrix}\mathbf{U}_q\mathbf{D}_q(\beta)\\ -\mathbf{U}_q\mathbf{K}_q(\beta)\end{bmatrix}). \tag{7}$$

In addition, $\mathbf{\Omega}$ has the following closed-form expression:

Lemma 3 $\mathbf{\Omega} = -(1+\beta)\begin{bmatrix}\mathbf{0}_{q\times(n-q)} & \mathbf{0}_{q\times(n-q)} & \mathrm{diag}(\frac{1}{\sqrt{\lambda_1^2-4\beta}}, \cdots, \frac{1}{\sqrt{\lambda_q^2-4\beta}})\end{bmatrix}$ *and* $\|\mathbf{\Omega}\|_2 = \frac{1+\beta}{\sqrt{\lambda_q^2-4\beta}}$.

Plugging Eq. (4) and Eq. (6) into Eq.(3) gives us that

$$\mathbf{Y}_{t+1}\widehat{\mathbf{R}}_{t+1} = \begin{bmatrix}\mathbf{V}_q & \mathbf{V}_{-q}\end{bmatrix}\begin{bmatrix}\mathbf{I} & \mathbf{\Omega}\\ \mathbf{0} & \mathbf{I}\end{bmatrix}\begin{bmatrix}\mathbf{\Lambda}_q & \mathbf{0}\\ \mathbf{0} & \mathbf{\Lambda}_{-q}\end{bmatrix}\begin{bmatrix}\mathbf{I} & \mathbf{\Omega}\\ \mathbf{0} & \mathbf{I}\end{bmatrix}^{-1}\begin{bmatrix}\mathbf{V}_q & \mathbf{V}_{-q}\end{bmatrix}^{\mathrm{H}}\mathbf{Y}_t + \mathbf{E}_t.$$

Pre-multiplying both sides of the above equation by

$$\begin{bmatrix}\mathbf{I} & \mathbf{\Omega}\\ \mathbf{0} & \mathbf{I}\end{bmatrix}^{-1}\begin{bmatrix}\mathbf{V}_q & \mathbf{V}_{-q}\end{bmatrix}^{\mathrm{H}} = \begin{bmatrix}\mathbf{I} & -\mathbf{\Omega}\\ \mathbf{0} & \mathbf{I}\end{bmatrix}\begin{bmatrix}\mathbf{V}_q^{\mathrm{H}}\\ \mathbf{V}_{-q}^{\mathrm{H}}\end{bmatrix} = \begin{bmatrix}\mathbf{V}_q^{\mathrm{H}} - \mathbf{\Omega}\mathbf{V}_{-q}^{\mathrm{H}}\\ \mathbf{V}_{-q}^{\mathrm{H}}\end{bmatrix} = \begin{bmatrix}\widehat{\mathbf{W}}_q^{\mathrm{H}}\\ \mathbf{V}_{-q}^{\mathrm{H}}\end{bmatrix},$$

we get that

$$\begin{bmatrix}\widehat{\mathbf{W}}_q^{\mathrm{H}}\\ \mathbf{V}_{-q}^{\mathrm{H}}\end{bmatrix}\mathbf{Y}_{t+1}\widehat{\mathbf{R}}_{t+1} = \begin{bmatrix}\mathbf{\Lambda}_q & \mathbf{0}\\ \mathbf{0} & \mathbf{\Lambda}_{-q}\end{bmatrix}\begin{bmatrix}\widehat{\mathbf{W}}_q^{\mathrm{H}}\\ \mathbf{V}_{-q}^{\mathrm{H}}\end{bmatrix}\mathbf{Y}_t + \begin{bmatrix}\widehat{\mathbf{W}}_q^{\mathrm{H}}\\ \mathbf{V}_{-q}^{\mathrm{H}}\end{bmatrix}\mathbf{E}_t,$$

i.e.,

$$\begin{cases}\widehat{\mathbf{W}}_q^{\mathrm{H}}\mathbf{Y}_{t+1}\widehat{\mathbf{R}}_{t+1} = \mathbf{\Lambda}_q\widehat{\mathbf{W}}_q^{\mathrm{H}}\mathbf{Y}_t + \widehat{\mathbf{W}}_q^{\mathrm{H}}\mathbf{E}_t,\\ \mathbf{V}_{-q}^{\mathrm{H}}\mathbf{Y}_{t+1}\widehat{\mathbf{R}}_{t+1} = \mathbf{\Lambda}_{-q}\mathbf{V}_{-q}^{\mathrm{H}}\mathbf{Y}_t + \mathbf{V}_{-q}^{\mathrm{H}}\mathbf{E}_t.\end{cases} \tag{8}$$

To make the non-diagonal (upper triangular) matrix $\mathbf{\Lambda}_{-q}$ amenable to the analysis in the sequel, we adapt the trick used in Lemma 7.3.2 of Golub and Van Loan (2013) to the noisy setting. First, there exists one permutation matrix $\mathbf{Q}$ such that

$$\mathbf{Q}\mathbf{\Lambda}_{-q}\mathbf{Q}^\top = \mathrm{diag}(\begin{bmatrix}\lambda_{q+1}^+ & -(1+(\lambda_{q+1}^-)^2)\\ 0 & \lambda_{q+1}^-\end{bmatrix}, \cdots, \begin{bmatrix}\lambda_n^+ & -(1+(\lambda_n^-)^2)\\ 0 & \lambda_n^-\end{bmatrix}, \lambda_1^-, \cdots, \lambda_q^-),$$

where, if $\lambda_s^+ = \sqrt{\beta}$ for $q < s \leq n$ then the corresponding 2×2 upper triangular matrix is replaced with $\begin{bmatrix} \sqrt{\beta} & 1+\beta \\ 0 & \sqrt{\beta} \end{bmatrix}$. Let $\mathbf{\Gamma} = \mathrm{diag}(\underbrace{\mathrm{diag}(1, 1+\gamma)}_{s=q+1}, \cdots, \underbrace{\mathrm{diag}(1, 1+\gamma)}_{s=n}, \mathbf{I}_q)$ with $1 + \gamma = \frac{16(1+\beta)}{(\lambda_k^+ - \sqrt{\beta})}$. We then have that

$$\mathbf{\Gamma Q \Lambda}_{-q} \mathbf{Q}^\top \mathbf{\Gamma}^{-1} = \mathrm{diag}(\begin{bmatrix} \lambda_{q+1}^+ & -\frac{1+(\lambda_{q+1}^-)^2}{1+\gamma} \\ 0 & \lambda_{q+1}^- \end{bmatrix}, \cdots, \begin{bmatrix} \lambda_n^+ & -\frac{1+(\lambda_n^-)^2}{1+\gamma} \\ 0 & \lambda_n^- \end{bmatrix}, \lambda_1^-, \cdots, \lambda_q^-). \quad (9)$$

Pre-multiplying both sides of the upper equation by $(\mathbf{I} + \mathbf{\Omega\Omega}^{\mathrm{H}})^{-\frac{1}{2}}$ and the lower equation by $\mathbf{\Gamma Q}$ in Eq.(8), we get that

$$\begin{cases} \mathbf{W}_q^{\mathrm{H}} \mathbf{Y}_{t+1} \widehat{\mathbf{R}}_{t+1} = \mathbf{\Lambda}_q \mathbf{W}_q^{\mathrm{H}} \mathbf{Y}_t + \mathbf{W}_q^{\mathrm{H}} \mathbf{E}_t \\ \mathbf{\Gamma Q V}_{-q}^{\mathrm{H}} \mathbf{Y}_{t+1} \widehat{\mathbf{R}}_{t+1} = \mathbf{\Gamma Q \Lambda}_{-q} \mathbf{V}_{-q}^{\mathrm{H}} \mathbf{Y}_t + \mathbf{\Gamma Q V}_{-q}^{\mathrm{H}} \mathbf{E}_t \end{cases}, \quad (10)$$

where the upper equation has used that $(\mathbf{I}+\mathbf{\Omega\Omega}^{\mathrm{H}})^{-\frac{1}{2}} \mathbf{\Lambda}_q = \mathbf{\Lambda}_q (\mathbf{I}+\mathbf{\Omega\Omega}^{\mathrm{H}})^{-\frac{1}{2}}$ because $(\mathbf{I}+\mathbf{\Omega\Omega}^{\mathrm{H}})^{-\frac{1}{2}}$ is diagonal by Lemma 3.

We are now in a position to define our potential function Φ_t and show its geometric shrinkage across iterations, in order to establish the convergence rate of Algorithm 1. Previous potential functions consider characterizations between $\mathbf{U}_k$ or $\mathbf{U}_q$ and $\mathbf{X}_t$ (i.e., subspaces of $\mathbb{R}^n$), while we need to consider a characterization between $\mathbf{V}_q$ and $\mathbf{Y}_t$ (i.e., subspaces of $\mathbb{R}^{2n}$) instead. Hardt and Price (2014) used as the potential function the tangent of the largest principal angle between $\mathbf{U}_k$ and $\mathbf{X}_t$: $\tan\theta_{\max}(\mathbf{U}_k, \mathbf{X}_t) = \|(\mathbf{U}_{-k}^\top \mathbf{X}_t)(\mathbf{U}_k^\top \mathbf{X}_t)^\dagger\|_2$, which admits geometric shrinkage. However, the geometric shrinkage might not hold under a higher level of noise (Hardt and Price, 2014) which can be allowed by using a larger iteration rank p than the target rank k. Gu (2015) considered $\|\mathbf{\Sigma}_{-q}^t (\mathbf{U}_{-q}^\top \mathbf{X}_0)(\mathbf{U}_q^\top \mathbf{X}_0)^\dagger \begin{bmatrix} \mathbf{\Sigma}_k^{-t} \\ \mathbf{0} \end{bmatrix} \|_2$ as the potential function for analyzing the noiseless power method and demonstrated the improved spectral gap $(\lambda_k - \lambda_{q+1})$ dependence of the convergence. Since this function can't handle noises across iterations, Balcan et al. (2016) proposed a variant, i.e., $h_t = \|(\mathbf{U}_{-q}^\top \mathbf{X}_t)(\mathbf{U}_q^\top \mathbf{X}_t)^\dagger \begin{bmatrix} \mathbf{I}_k \\ \mathbf{0} \end{bmatrix} \|_2$, to adapt the analysis to the presence of the noises per iteration. To serve our purpose, we further propose the following calibrated variant:

$$\Phi_t = \|(\mathbf{\Gamma Q V}_{-q}^{\mathrm{H}} \mathbf{Y}_t)(\mathbf{W}_q^{\mathrm{H}} \mathbf{Y}_t)^\dagger \begin{bmatrix} \mathbf{I}_k \\ \mathbf{0} \end{bmatrix} \|_2.$$

We call our potential function Φ_t the calibrated rank-k perturbation on $\mathbf{V}_q$ by $\mathbf{Y}_t$, and have the following key lemma about Φ_t.

Lemma 4 *Under the noise conditions given in Theorem 1 on $\boldsymbol{\xi}_t$, it holds that*

$$\Phi_{t+1} \leq (1 - \frac{1}{4}\sqrt{\frac{\lambda_k - 2\sqrt{\beta}}{\lambda_k}}) \Phi_t.$$

Note that the geometric shrinkage of the potential function $g_t = h_t - \frac{C\epsilon}{1-C\epsilon}$ (for a sufficiently small constant $0 < C < 1$) in Balcan et al. (2016) does not hold under our noise conditions, because

the $O(\epsilon)$ term in g_t will inevitably require the norm of the noise term to be as small as $O(\epsilon)$ across iterations (see the proof of their Lemma 2.1 or A.1). Moreover, if we follow the analysis of Balcan et al. (2016) including the use of $(\Phi_t - \frac{C\epsilon}{1-C\epsilon})$ as our potential function, it will end up with an even worse spectral gap dependence $O((\lambda_k - \lambda_{q+1})^2\epsilon)$ of the noise tolerance. On the other hand, the noise tolerance of Balcan et al. (2016) can be improved similarly if our analysis technique is applied there (see the proof of our Lemma 4 in Appendix).

To state the convergence results of Algorithm 1 in terms of $\sin\theta_{\max}(\mathbf{X}_t, \mathbf{U}_k)$, we can relate $\sin\theta_{\max}(\mathbf{X}_t, \mathbf{U}_k)$ to Φ_t via $\sin\theta_{\max}(\mathbf{Y}_t, \mathbf{V}_k)$, based on the following two lemmas about the relation between Φ_t and $\sin\theta_{\max}(\mathbf{Y}_t, \mathbf{V}_k)$ as well as the relation between $\sin\theta_{\max}(\mathbf{Y}_t, \mathbf{V}_k)$ and $\sin\theta_{\max}(\mathbf{X}_t, \mathbf{U}_k)$, respectively. Here $\mathbf{V}_k$ represents the dominant k-dimensional invariant subspace of $\mathbf{B}$ spanned by the first k columns of $\mathbf{V}_q$.

Lemma 5 $\sin\theta_{\max}(\mathbf{Y}_t, \mathbf{V}_k) = \|(\mathbf{I} - \mathbf{Y}_t\mathbf{Y}_t^\top)\mathbf{V}_k\|_2 \le (1 + 2\frac{1+\beta}{\sqrt{\lambda_q^2 - 4\beta}})\Phi_t$.

Lemma 6 *If* $\sin\theta_{\max}(\mathbf{Y}_t, \mathbf{V}_k) < \frac{\lambda_k^+}{k\sqrt{1+(\lambda_k^+)^2}}\epsilon$, *then* $\sin\theta_{\max}(\mathbf{X}_t, \mathbf{U}_k) < \epsilon$.

By Lemma 4, it holds that $\Phi_T \le (1 - \frac{1}{4}\sqrt{\frac{\lambda_k - 2\sqrt{\beta}}{\lambda_k}})^T\Phi_0 < \exp\{-\frac{T}{4}\sqrt{\frac{\lambda_k - 2\sqrt{\beta}}{\lambda_k}}\}\Phi_0$. In order to have that $\sin\theta_{\max}(\mathbf{X}_T, \mathbf{U}_k) < \epsilon$, by Lemma 5 and Lemma 6, it suffices to make T satisfy that

$$(1 + 2\frac{1+\beta}{\sqrt{\lambda_q^2 - 4\beta}})\exp\{-\frac{T}{4}\sqrt{\frac{\lambda_k - 2\sqrt{\beta}}{\lambda_k}}\}\Phi_0 \le \frac{\lambda_k^+}{k\sqrt{1+(\lambda_k^+)^2}}\epsilon.$$

Thus, we have $T \ge \widehat{T} = 4\sqrt{\frac{\lambda_k}{\lambda_k - 2\sqrt{\beta}}}\log(2k\sqrt{\frac{1+(\lambda_k^+)^2}{(\lambda_k^+)^2}}\frac{1+\beta+\lambda_q}{\sqrt{\lambda_q^2-4\beta}}\frac{\Phi_0}{\epsilon})$. Moreover,

$$\begin{aligned}
\log\Phi_0 &= \log\|\mathbf{\Gamma}\mathbf{Q}\mathbf{V}_{-q}^{\mathrm{H}}\begin{bmatrix}\widehat{\mathbf{X}}_0\\ \mathbf{0}\end{bmatrix}(\mathbf{O}^{\mathrm{H}}\begin{bmatrix}\mathbf{U}_q\mathbf{D}_q(\beta)\\ -\mathbf{U}_q\mathbf{K}_q(\beta)\end{bmatrix}^{\mathrm{H}}\begin{bmatrix}\widehat{\mathbf{X}}_0\\ \mathbf{0}\end{bmatrix})^\dagger\begin{bmatrix}\mathbf{I}_k\\ \mathbf{0}\end{bmatrix}\|_2 \\
&\le \log(\|\mathbf{\Gamma}\|_2\|\mathbf{Q}\|_2) + \log(\|\mathbf{V}_{-q}^{\mathrm{H}}\|_2\|\widehat{\mathbf{X}}_0\|_2) + \log(\|(\mathbf{O}\mathbf{D}_q(\beta)\mathbf{U}_q^\top\widehat{\mathbf{X}}_0)^\dagger\|_2\|\begin{bmatrix}\mathbf{I}_k\\ \mathbf{0}\end{bmatrix}\|_2) \\
&\le \log(1+\gamma) + \log\|\widehat{\mathbf{X}}_0\|_2 - \log\sigma_{\min}(\mathbf{D}_q(\beta)\mathbf{U}_q^\top\widehat{\mathbf{X}}_0) \\
&\le \log(1+\gamma) - \log\sigma_{\min}(\mathbf{D}_q(\beta)) + \log\frac{\|\widehat{\mathbf{X}}_0\|_2}{\sigma_{\min}(\mathbf{U}_q^\top\widehat{\mathbf{X}}_0)} \\
&\le \log\frac{16(1+\beta)}{\lambda_k^+ - \sqrt{\beta}} + \log\sqrt{\frac{\beta^2 + (\lambda_k^+)^2}{(\lambda_k^+)^2}} + \log\frac{\tau\sqrt{n}}{\sqrt{p} - \sqrt{q-1}},
\end{aligned}$$

where the first equality ($\mathbf{R}_0$ in $\mathbf{Y}_0$ has been cancelled there) has used that $\mathbf{W}_q = \begin{bmatrix}\mathbf{U}_q\mathbf{D}_q(\beta)\\ -\mathbf{U}_q\mathbf{K}_q(\beta)\end{bmatrix}\mathbf{O}$ for a certain unitary matrix $\mathbf{O} \in \mathbb{C}^{q\times q}$ by Eq. (7), the second last inequality has used the inequality that $\sigma_{\min}(\mathbf{D}_q(\beta)\mathbf{U}_q^\top\widehat{\mathbf{X}}_0) \ge \sigma_{\min}(\mathbf{D}_q(\beta))\sigma_{\min}(\mathbf{U}_q^\top\widehat{\mathbf{X}}_0)$, and the last inequality has used Lemma 2 and that $\frac{\|\widehat{\mathbf{X}}_0\|_2}{\sigma_{\min}(\mathbf{U}_q^\top\widehat{\mathbf{X}}_0)} \le \frac{\tau\sqrt{n}}{\sqrt{p}-\sqrt{q-1}}$ with probability at least $1 - \tau^{-\Omega(p+1-q)} - e^{-\Omega(n)}$ by Lemma 2.5

in Hardt and Price (2014). Hence,

$$\widehat{T} \leq 4\sqrt{\frac{\lambda_k}{\lambda_k - 2\sqrt{\beta}}}\log(\sqrt{\frac{1+(\lambda_k^+)^2}{(\lambda_k^+)^2}\frac{\beta^2+(\lambda_k^+)^2}{(\lambda_k^+)^2}}\frac{1+\beta+\lambda_q}{\sqrt{\lambda_q^2-4\beta}}\frac{1+\beta}{\lambda_k^+-\sqrt{\beta}}\frac{32k\tau\sqrt{n}}{\sqrt{p}-\sqrt{q-1}}\frac{1}{\epsilon}) \triangleq \widetilde{T}.$$

Therefore, $\sin\theta_{\max}(\mathbf{X}_t, \mathbf{U}_k) < \epsilon$ is guaranteed by $T \geq \widetilde{T}$ with high probability, which completes the proof of Theorem 1.

Remark 5 The proofs of the results under other values of β mentioned in Remark 3 are straightforward, once we note the non-zero block shape changes of three matrices ($\mathbf{\Lambda}_q$, $\mathbf{\Lambda}_{q,-q}$, and $\mathbf{\Lambda}_{-q}$) in Eq. (4) and the consequent change of $\mathbf{\Omega}$ based on Lemma 2.

4. Applications

In this section, we propose faster distributed or streaming PCA (Hardt and Price, 2014; Liang et al., 2014; Balcan et al., 2016) algorithm as downstream applications of our Algorithm 1, and show their improved communication or sample complexity over the plain counterpart based on Theorem 1. It is worth noting that $\boldsymbol{\xi}_t = \mathbf{0}$ for the distributed setting[1] and $\boldsymbol{\xi}_t \neq \mathbf{0}$ for the streaming setting considered here.

4.1. Distributed PCA

We consider the distributed PCA model of $S \geq 1$ computing nodes and a central computing node. Each computing node j stores either a positive semi-definite matrix $\mathbf{A}^{(j)}$ or a set of sample data points with sample covariance matrix being $\mathbf{A}^{(j)}$, while the central node has no data stored. The goal is to estimate the top-k eigenspace $\mathbf{U}_k$ of the aggregated data matrix $\mathbf{A} = \sum_{j=1}^S \mathbf{A}^{(j)}$, under the constraint that there is only a public channel for communication between computing nodes and the central node. Liang et al. (2014) gave an $O(\frac{pnS}{\epsilon})$ communication complexity for this distributed PCA model, which was improved exponentially by Balcan et al. (2016). We further propose faster distributed PCA algorithm with pseudo codes given in Algorithm 2.

Theorem 7 *For $2\sqrt{\beta}$ close to λ_{q+1}, intermediate rank q satisfying $k \leq q \leq p$, and $T = O(\sqrt{\frac{\lambda_k}{\lambda_k-\lambda_{q+1}}}\log\frac{n}{(\lambda_q-\lambda_{q+1})\epsilon})$, Algorithm 2 outputs $\mathbf{X}_T$ such that $\sin\theta_{\max}(\mathbf{X}_T, \mathbf{U}_k) < \epsilon$ with high probability (w.h.p.) and communication complexity at $O(pnTS) = O(pnS\sqrt{\frac{\lambda_k}{\lambda_k-\lambda_{q+1}}}\log\frac{n}{(\lambda_q-\lambda_{q+1})\epsilon})$.*

The proof is straightforward as it is the simplest case of Theorem 1 with $\boldsymbol{\xi}_t = \mathbf{0}$. Theorem 7 significantly improves over the communication complexity bound $O(pnS\frac{\lambda_k}{\lambda_k-\lambda_{q+1}}\log\frac{n}{\epsilon})$ in Balcan et al. (2016). For this application, we can use varying β by setting $2\sqrt{\beta_t}$ to be the $(q+1)$-th largest diagonal entry of $\widehat{\mathbf{\Sigma}}_p = \mathbf{X}_t^\top \sum_{j=1}^S \mathbf{X}_t^{(j)} \in \mathbb{R}^{p\times p}$.

1. As opposed to Balcan et al. (2016), privacy is not considered here as it has not been clear yet if the noise condition in Eq. (2) can be met by the privacy constraint.

Algorithm 2 Faster distributed PCA

Input: Data matrices $\mathbf{A}^{(1)}, \cdots, \mathbf{A}^{(S)} \in \mathbb{R}^{n \times n}$ distributed over S computing nodes, momentum parameter $\beta > 0$, target rank k, iteration rank $p \geq k$, iteration number T.
Output: approximate top-k eigenspace $\mathbf{X}_T$.
The central node sets $\mathbf{X}_{-1} = \mathbf{0} \in \mathbb{R}^{n \times p}$ and QR factorize an entry-wise i.i.d. standard Gaussian matrix $\widehat{\mathbf{X}}_0 \in \mathbb{R}^{n \times p}$ such that $\mathbf{X}_0 \mathbf{R}_0 = \widehat{\mathbf{X}}_0$
for $t = 0, 1, \cdots, T-1$ **do**
    The central node broadcasts $\mathbf{X}_t$ to all S computing nodes
    for $j = 1, \cdots, S$ **do**
        Computing node j computes $\mathbf{X}_t^{(j)} = \mathbf{A}^{(j)} \mathbf{X}_t$ and send $\mathbf{X}_t^{(j)}$ back to the central node
    end for
    The central node computes $\widehat{\mathbf{X}}_{t+1} = \sum_{j=1}^{S} \mathbf{X}_t^{(j)} - \beta \mathbf{X}_{t-1} \mathbf{R}_t^{-1}$
    The central node QR factorizes $\widehat{\mathbf{X}}_{t+1}$ such that $\mathbf{X}_{t+1} \mathbf{R}_{t+1} = \widehat{\mathbf{X}}_{t+1}$
end for

Algorithm 3 Faster streaming PCA

Input: I.I.D. data stream $\mathbf{z}_1, \cdots, \mathbf{z}_m \sim \mathcal{D}$, momentum parameter $\beta > 0$, target rank k, iteration rank $p \geq k$, iteration number T.
Output: approximate top-k eigenspace $\mathbf{X}_T$.
Set $M = \lfloor \frac{m}{T} \rfloor$, $\mathbf{X}_{-1} = \mathbf{0} \in \mathbb{R}^{n \times p}$ and QR factorize an entry-wise i.i.d. standard Gaussian matrix $\widehat{\mathbf{X}}_0 \in \mathbb{R}^{n \times p}$ such that $\mathbf{X}_0 \mathbf{R}_0 = \widehat{\mathbf{X}}_0$
for $t = 0, 1, \cdots, T-1$ **do**
    $\widehat{\mathbf{X}}_{t+1} = \mathbf{A}_t \mathbf{X}_t - \beta \mathbf{X}_{t-1} \mathbf{R}_t^{-1}$ where $\mathbf{A}_t = \sum_{j=(t-1)M+1}^{tM} \mathbf{z}_j \mathbf{z}_j^\top$
    QR factorizes $\widehat{\mathbf{X}}_{t+1}$ such that $\mathbf{X}_{t+1} \mathbf{R}_{t+1} = \widehat{\mathbf{X}}_{t+1}$
end for

4.2. Memory-Efficient Streaming PCA

Given a stream of i.i.d. samples $\mathbf{z}_1, \cdots, \mathbf{z}_m \in \mathbb{R}^{n \times 1}$ drawn from an unknown distribution $\mathcal{D}$, the goal of the streaming PCA is to estimate the top-k eigenspace of the population covariance matrix $\mathbf{A} = \mathbf{E}_{\mathbf{z} \sim \mathcal{D}}[\mathbf{z}\mathbf{z}^\top] \in \mathbb{R}^{n \times n}$ with $O(kn)$ memory. Under the framework of the noisy power method, a natural algorithm was introduced for this problem with the spiking covariance model in Mitliagkas et al. (2013), while Hardt and Price (2014) analyzed this algorithm for a broader class of distributions (i.e., (B, p)-round distributions) that have fast tail decay based on their analysis of the noisy power method. Balcan et al. (2016) improved the results of Hardt and Price (2014) in terms of gap dependency. In this work, we further improve the results of Balcan et al. (2016) in terms of gap dependency by a faster streaming PCA algorithm described in Algorithm 3. There are other analyses on streaming PCA (Allen-Zhu and Li, 2017), but for a different setting where the iteration rank p and the target rank k are the same.

Definition 8 *((B,p)-round distributions,Hardt and Price (2014)) A distribution $\mathcal{D}$ over $\mathbb{R}^n$ is said to be (B,p)-round, if it holds for every p-dimension projection $\mathbf{\Pi}$ and all $t \geq 1$ that*

$$\max\{\ \Pr_{\mathbf{z}\sim\mathcal{D}}[\|\mathbf{z}\|_2 \geq t], \Pr_{\mathbf{z}\sim\mathcal{D}}[\|\mathbf{\Pi z}\|_2 \geq t\sqrt{\frac{Bp}{n}}]\ \} \leq \exp\{-t\}.$$

Theorem 9 *Suppose $\mathcal{D}$ is a (B,p)-round distribution. For $2\sqrt{\beta}$ close to λ_{q+1}, intermediate rank q satisfying $k \leq q \leq p$, and $T = O(\sqrt{\frac{\lambda_k}{\lambda_k-\lambda_{q+1}}}\log\frac{n}{(\lambda_q-\lambda_{q+1})\epsilon})$, Algorithm 3 outputs $\mathbf{X}_T$ such that $\sin\theta_{\max}(\mathbf{X}_T,\mathbf{U}_k) < \epsilon$ with probability at least 0.9 and sample complexity at $m = MT = \widetilde{\Omega}(\frac{B^2p\log^2 n}{\epsilon^2(\lambda_k-\lambda_{q+1})^2 n}\sqrt{\frac{\lambda_k}{\lambda_k-\lambda_{q+1}}})$, where $\widetilde{\Omega}(\cdot)$ hides logarithmic factors.*

Proof The proof is similar to that of Theorem 4.2 in Balcan et al. (2016). Setting $\boldsymbol{\xi}_t = (\mathbf{A}_t - \mathbf{A})\mathbf{X}_t$, then by Lemma 3.5 of Hardt and Price (2014), $M = \widetilde{\Omega}(\frac{B^2p\log^2 n}{\epsilon^2(\lambda_k-\lambda_{q+1})^2 n})$ suffices to guarantee that $\boldsymbol{\xi}_t$ satisfies the noise conditions in Theorem 1 w.h.p. as it satisfies conditions of Theorem 2.2 in Balcan et al. (2016) w.h.p. (see Table 1 and Remark 1). Thus, the total number of data points needed is $m = MT = \widetilde{\Omega}(\frac{B^2p\log^2 n}{\epsilon^2(\lambda_k-\lambda_{q+1})^2 n}\sqrt{\frac{\lambda_k}{\lambda_k-\lambda_{q+1}}})$. □

Theorem 9 improves over the sample complexity $\widetilde{\Omega}(\frac{\lambda_k B^2p\log^2 n}{\epsilon^2(\lambda_k-\lambda_{q+1})^3 n})$ given in Balcan et al. (2016). For this application, we may use varying β by setting $2\sqrt{\beta_t}$ to be the $(q+1)$-th largest diagonal entry of $\widehat{\mathbf{\Sigma}}_p = \frac{1}{tM}\sum_{t'=1}^{t}\mathbf{X}_{t'}^\top\mathbf{A}_{t'}\mathbf{X}_{t'} \in \mathbb{R}^{p\times p}$.

5. Conclusion

In this work, we presented faster noisy power method and gave a novel analysis to show further improved spectral gap dependency over state-of-the-art results for the noisy power method by Balcan et al. (2016). Our Algorithm 1 can serve as a faster meta algorithm with applications to downstream tasks such as distributed PCA and streaming PCA, and have their theoretical guarantees improved in terms of spectral gap dependence. Limitations of our analysis lie at its inherent gap-dependency, due to the condition that $\lambda_q > \lambda_{q+1}$ though it is not the harsher condition that $\lambda_k > \lambda_{k+1}$, and additionally logarithmic gap dependence, i.e., $\log\frac{1}{\lambda_q-\lambda_{q+1}}$. Particularly, it remains unknown if the spectral gap dependency of the noise tolerance can be improved to be the same as that of the convergence rate in terms of $\|\boldsymbol{\xi}_t\|_2$. These may be the artifacts of our analysis technique, and it will be interesting to consider removing them in future.

Acknowledgments

We would like to thank the anonymous reviewers for their helpful comments.

References

Zeyuan Allen-Zhu and Yuanzhi Li. First efficient convergence for streaming k-pca: A global, gap-free, and near-optimal rate. In *Proceedings of the 58th IEEE Annual Symposium on Foundations of Computer Science (FOCS)*, pages 487–492, Berkeley, CA, 2017.

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.

Rajendra Bhatia. *Matrix Analysis*, volume 169. Springer, 1997.

G.H. Golub and C.F. Van Loan. *Matrix Computations*. Johns Hopkins Studies in the Mathematical Sciences. Johns Hopkins University Press, 2013.

Ming Gu. Subspace iteration randomization and singular value problems. *SIAM J. Sci. Comput.*, 37 (3), 2015.

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

Moritz Hardt and Aaron Roth. Beyond worst-case analysis in private singular vector computation. In *Proceedings of the Symposium on Theory of Computing Conference (STOC)*, pages 331–340, Palo Alto, CA, 2013.

Yingyu Liang, Maria-Florina Balcan, Vandana Kanchanapally, and David P. Woodruff. Improved distributed principal component analysis. In *Advances in Neural Information Processing Systems (NIPS)*, pages 3113–3121, Montreal, Canada, 2014.

Ziqi Liu, Yu-Xiang Wang, and Alexander J. Smola. Fast differentially private matrix factorization. In *Proceedings of the 9th ACM Conference on Recommender Systems (RecSys)*, pages 171–178, Vienna, Austria, 2015.

Ioannis Mitliagkas, Constantine Caramanis, and Prateek Jain. Memory limited, streaming PCA. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2886–2894, Lake Tahoe, NV, 2013.

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.

Shusen Wang, Zhihua Zhang, and Tong Zhang. Improved analyses of the randomized power method and block lanczos method. *CoRR*, abs/1508.06429, 2015.

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, 2018.

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

Zhiqiang Xu and Ping Li. A practical riemannian algorithm for computing dominant generalized eigenspace. In *Proceedings of the Thirty-Sixth Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 819–828, virtual online, 2020.

Zhiqiang Xu and Ping Li. On the riemannian search for eigenvector computation. *Journal of Machine Learning Research*, 22(249):1–46, 2021.

Lemma 2.1 If there exists an integer s such that $\lambda_s > 2\sqrt{\beta} \geq \lambda_{s+1}$ and $\lambda_n > 0$, we then have the following Schur decomposition for $\mathbf{B} = \begin{bmatrix} \mathbf{A} & -\beta\mathbf{I} \\ \mathbf{I} & \mathbf{0} \end{bmatrix}$:

$$\begin{bmatrix} \mathbf{A} & -\beta\mathbf{I} \\ \mathbf{I} & \mathbf{0} \end{bmatrix} = \begin{bmatrix} \mathbf{U}_s\mathbf{D}_s(1) & \mathbf{U}_{-s}\mathbf{D}_{-s}(1) & -\mathbf{U}_{-s}\mathbf{K}_{-s}(1) & -\mathbf{U}_s\mathbf{K}_s(1) \\ \mathbf{U}_s\mathbf{K}_s(1) & \mathbf{U}_{-s}\mathbf{K}_{-s}(1) & \mathbf{U}_{-s}\overline{\mathbf{D}_{-s}(1)} & \mathbf{U}_s\overline{\mathbf{D}_s(1)} \end{bmatrix} \times$$
$$\begin{bmatrix} \mathbf{\Sigma}_s^+ & \mathbf{0} & \mathbf{0} & -(1+\beta)\mathbf{I} \\ \mathbf{0} & \mathbf{\Sigma}_{-s}^+ & \mathbf{T}(-1) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{\Sigma}_{-s}^- & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{\Sigma}_s^- \end{bmatrix} \begin{bmatrix} \mathbf{U}_s\mathbf{D}_s(1) & \mathbf{U}_{-s}\mathbf{D}_{-s}(1) & -\mathbf{U}_{-s}\mathbf{K}_{-s}(1) & -\mathbf{U}_s\mathbf{K}_s(1) \\ \mathbf{U}_s\mathbf{K}_s(1) & \mathbf{U}_{-s}\mathbf{K}_{-s}(1) & \mathbf{U}_{-s}\overline{\mathbf{D}_{-s}(1)} & \mathbf{U}_s\overline{\mathbf{D}_s(1)} \end{bmatrix}^{\mathrm{H}},$$

where

$$\lambda_j^{\pm} = \frac{\lambda_j \pm \sqrt{\lambda_j^2 - 4\beta}}{2}, \qquad \mathbf{\Sigma}_s^+ = \mathrm{diag}(\lambda_1^+, \cdots, \lambda_s^+), \qquad \mathbf{\Sigma}_{-s}^+ = \mathrm{diag}(\lambda_{s+1}^+, \cdots, \lambda_n^+),$$
$$\mathbf{\Sigma}_s^- = \mathrm{diag}(\lambda_1^-, \cdots, \lambda_s^-), \qquad \mathbf{\Sigma}_{-s}^- = \mathrm{diag}(\lambda_{s+1}^-, \cdots, \lambda_n^-),$$
$$\mathbf{D}(a) = \mathrm{diag}(\tfrac{\lambda_1^+}{\sqrt{a^2+|\lambda_1^+|^2}}, \cdots, \tfrac{\lambda_s^+}{\sqrt{a^2+|\lambda_s^+|^2}}), \qquad \mathbf{D}_{-s}(a) = \mathrm{diag}(\tfrac{\lambda_{s+1}^+}{\sqrt{a^2+|\lambda_{s+1}^+|^2}}, \cdots, \tfrac{\lambda_n^+}{\sqrt{a^2+|\lambda_n^+|^2}}),$$
$$\mathbf{K}(a) = \mathrm{diag}(\tfrac{a}{\sqrt{a^2+|\lambda_1^+|^2}}, \cdots, \tfrac{a}{\sqrt{a^2+|\lambda_s^+|^2}}), \qquad \mathbf{K}_{-s}(a) = \mathrm{diag}(\tfrac{a}{\sqrt{a^2+|\lambda_{s+1}^+|^2}}, \cdots, \tfrac{a}{\sqrt{a^2+|\lambda_n^+|^2}}),$$

and $\mathbf{T}(b) = b\mathbf{I} + \frac{1}{b}(\mathbf{\Sigma}_{-s}^-)^2$. In addition, if there exists any $j > s$ such that $\lambda_j = 2\sqrt{\beta}$, then $\lambda_j^{\pm} = \sqrt{\beta}$ and the corresponding diagonal entry in the block $\mathbf{T}(-1)$ is replaced with $(1+\beta)$.

Proof The full eigenvalue decomposition of the given real symmetric matrix is

$$\mathbf{A} = \begin{bmatrix} \mathbf{U}_s & \mathbf{U}_{-s} \end{bmatrix} \mathrm{diag}(\mathbf{\Sigma}_s, \mathbf{\Sigma}_{-s}) \begin{bmatrix} \mathbf{U}_s & \mathbf{U}_{-s} \end{bmatrix}^{\top}.$$

See Section 2 in the main paper for notations. Suppose one eigenpair of $\mathbf{A}$ is $(\lambda_j, \mathbf{u}_j)$. By Proposition 9 in Xu et al. (2018), a double eigenpair of $\mathbf{B}$ then can be written as $\left(\frac{\lambda_j \pm \sqrt{\lambda_j^2 - 4\beta}}{2}, \begin{bmatrix} \frac{\lambda_j \pm \sqrt{\lambda_j^2 - 4\beta}}{2}\mathbf{u}_j \\ \mathbf{u}_j \end{bmatrix} \right)$. For $\lambda_s > 2\sqrt{\beta} \geq \lambda_{s+1}$, all the eigenvectors of $\mathbf{B}$ in the above double eigenpairs together (i.e., $j = 1, \cdots, n$) will span the whole Euclidean complex vector space $\mathbb{C}^{2n}$ as long as there exists no $j > s$ such that $\lambda_j = 2\sqrt{\beta}$, otherwise the double eigenpair collapses into a single one for each j of the kind. In fact, $\mathbf{B}$ has the following characteristic polynomial:

$$\begin{aligned} & \det(\mathbf{B} - \mu\mathbf{I}) \\ = \ & \det \begin{bmatrix} \mathbf{A} - \mu\mathbf{I} & -\beta\mathbf{I} \\ \mathbf{I} & -\mu\mathbf{I} \end{bmatrix} = \det \begin{bmatrix} \mathrm{diag}(\mathbf{\Sigma}_s, \mathbf{\Sigma}_{-s}) - \mu\mathbf{I} & -\beta\mathbf{I} \\ \mathbf{I} & -\mu\mathbf{I} \end{bmatrix} \\ = \ & \det(\mathrm{diag}(\begin{bmatrix} \lambda_1 - \mu & -\beta \\ 1 & -\mu \end{bmatrix}, \cdots, \begin{bmatrix} \lambda_s - \mu & -\beta \\ 1 & -\mu \end{bmatrix}, \begin{bmatrix} \lambda_{s+1} - \mu & -\beta \\ 1 & -\mu \end{bmatrix}, \cdots, \begin{bmatrix} \lambda_n - \mu & -\beta \\ 1 & -\mu \end{bmatrix})) \\ = \ & \det \begin{bmatrix} \lambda_1 - \mu & -\beta \\ 1 & -\mu \end{bmatrix} \cdots \det \begin{bmatrix} \lambda_s - \mu & -\beta \\ 1 & -\mu \end{bmatrix} \det \begin{bmatrix} \lambda_{s+1} - \mu & -\beta \\ 1 & -\mu \end{bmatrix} \cdots \det \begin{bmatrix} \lambda_n - \mu & -\beta \\ 1 & -\mu \end{bmatrix}. \end{aligned}$$

Thus, $\mathbf{B}$'s eigenvalue $\mu = \sqrt{\beta}$ has algebraic multiplicity $2m$ corresponding to $\mathbf{A}$'s eigenvalue $2\sqrt{\beta}$ of multiplicity m, i.e., $\lambda_j = \cdots = \lambda_{j-m+1} = 2\sqrt{\beta} > \lambda_{j-m}$, when it happens. Meanwhile, the

corresponding geometric multiplicity is as follows:

$$\begin{aligned}\dim(\ker(\mathbf{B}-\sqrt{\beta}\mathbf{I})) &= 2n-\dim(\mathrm{col}(\mathbf{B}-\sqrt{\beta}\mathbf{I}))=2n-\mathrm{rank}(\mathbf{B}-\sqrt{\beta}\mathbf{I})\\ &= 2n-\sum_{j=1}^{n}\mathrm{rank}\begin{bmatrix}\lambda_j-\sqrt{\beta} & -\beta\\ 1 & -\sqrt{\beta}\end{bmatrix}=m,\end{aligned}$$

where $\dim(\cdot)$ represents the dimensionality of a space, $\ker(\square)$ represents the kernel space of a matrix, and $\mathrm{col}(\square)$ represents the column space of a matrix. This shows that, in addition to the collapsed single eigenpair, for each j of the kind, we need to derive one generalized eigenvector $\mathbf{z}_j\in\mathbb{R}^{2n\times 1}$ of $\mathbf{B}$ from the collapsed single eigenpair to span $\mathbb{C}^{2n}$, i.e., $(\mathbf{B}-\sqrt{\beta}\mathbf{I})\mathbf{z}_j=\begin{bmatrix}\sqrt{\beta}\mathbf{u}_j\\ \mathbf{u}_j\end{bmatrix}$. It is easy to see that $\mathbf{z}_j=\begin{bmatrix}\mathbf{u}_j\\ \mathbf{0}\end{bmatrix}$ which keeps orthogonal to all the eigenvectors or generalized eigenvectors of $\mathbf{B}$ corresponding to $\lambda_{j'}$ for $j'\neq j$. We then can list $\mathbf{B}$'s eigenvalues as well as corresponding eigenvectors or generalized eigenvectors in descending order of eigenvalues' magnitude in Table 3 where, if there is any $j>s$ such that $\lambda_j=2\sqrt{\beta}$ then the corresponding row in the middle is replaced by the following one:

$$|\lambda_j^{\pm}|=\sqrt{\beta},\quad \lambda_j^{\pm}\triangleq\sqrt{\beta},\quad \mathbf{u}_j^{+}\triangleq\begin{bmatrix}\sqrt{\beta}\mathbf{u}_j\\ \mathbf{u}_j\end{bmatrix},\quad \mathbf{u}_j^{-}\triangleq\begin{bmatrix}\mathbf{u}_j\\ \mathbf{0}\end{bmatrix}.$$

Thus, we have $\mathbf{B}\mathbf{u}_j^{\pm}=\lambda_j^{\pm}\mathbf{u}_j^{\pm}$ for j such that $\lambda_j\neq 2\sqrt{\beta}$, otherwise $\begin{cases}\mathbf{B}\mathbf{u}_j^{+}=\lambda_j^{+}\mathbf{u}_j^{+},\\ \mathbf{B}\mathbf{u}_j^{-}=\mathbf{u}_j^{+}+\lambda_j^{-}\mathbf{u}_j^{-}.\end{cases}$ In matrix form, it can be written as $\mathbf{B}\widehat{\mathbf{V}}_{\boxplus}=\widehat{\mathbf{V}}_{\boxplus}\widehat{\mathbf{\Lambda}}_{\boxplus}$, where

$$\widehat{\mathbf{V}}_{\boxplus}=\begin{bmatrix}\mathbf{u}_1^{+} & \mathbf{u}_1^{-} & \cdots & \mathbf{u}_n^{+} & \mathbf{u}_n^{-}\end{bmatrix}\quad,\quad \widehat{\mathbf{\Lambda}}_{\boxplus}=\mathrm{diag}(\widehat{\mathbf{\Lambda}}_1,\cdots,\widehat{\mathbf{\Lambda}}_n)\quad,\quad \widehat{\mathbf{\Lambda}}_j=\mathrm{diag}(\lambda_j^{+},\lambda_j^{-})$$

where for all j such that $\lambda_j=2\sqrt{\beta}$, each of the corresponding 2×2 diagonal blocks is replaced with the 2×2 upper triangular block $\widehat{\mathbf{\Lambda}}_j=\begin{bmatrix}\lambda_j^{+} & 1\\ 0 & \lambda_j^{-}\end{bmatrix}$. In order to derive $\mathbf{B}$'s Schur decomposition, we need to orthonormalize all the (generalized) eigenvectors of $\mathbf{B}$. For this purpose, in fact, we only need to orthonormalize the pair $(\mathbf{u}_j^{+},\mathbf{u}_j^{-})$ for each $j=1,\cdots,n$, because $(\mathbf{u}_j^{\pm})^{\mathrm{H}}\mathbf{u}_{j'}^{\pm}=0$ for any $j\neq j'$. There are two cases. If $\lambda_j\neq 2\sqrt{\beta}$, then we have the orthonormalization:

$$\begin{bmatrix}\mathbf{u}_j^{+} & \mathbf{u}_j^{-}\end{bmatrix}=\begin{bmatrix}\lambda_j^{+}\mathbf{u}_j & \lambda_j^{-}\mathbf{u}_j\\ \mathbf{u}_j & \mathbf{u}_j\end{bmatrix}=\underbrace{\begin{bmatrix}\frac{\lambda_j^{+}}{\sqrt{1+|\lambda_j^{+}|^2}}\mathbf{u}_j & -\frac{1}{\sqrt{1+|\lambda_j^{+}|^2}}\mathbf{u}_j\\ \frac{1}{\sqrt{1+|\lambda_j^{+}|^2}}\mathbf{u}_j & \frac{\overline{\lambda_j^{+}}}{\sqrt{1+|\lambda_j^{+}|^2}}\mathbf{u}_j\end{bmatrix}}_{\triangleq\ (\mathbf{v}_{2j-1},\mathbf{v}_{2j})}\underbrace{\begin{bmatrix}\sqrt{1+|\lambda_j^{+}|^2} & \frac{1+\overline{\lambda_j^{+}}\lambda_j^{-}}{\sqrt{1+|\lambda_j^{+}|^2}}\\ 0 & \frac{\lambda_j^{+}-\lambda_j^{-}}{\sqrt{1+|\lambda_j^{+}|^2}}\end{bmatrix}}_{\triangleq\ \mathbf{S}_j}.$$

If $\lambda_j=2\sqrt{\beta}$ then the orthonormalization is

$$\begin{bmatrix}\mathbf{u}_j^{+} & \mathbf{u}_j^{-}\end{bmatrix}=\begin{bmatrix}\sqrt{\beta}\mathbf{u}_j & \mathbf{u}_j\\ \mathbf{u}_j & \mathbf{0}\end{bmatrix}=\underbrace{\begin{bmatrix}\sqrt{\frac{\beta}{1+\beta}}\mathbf{u}_j & \frac{1}{\sqrt{1+\beta}}\mathbf{u}_j\\ \frac{1}{\sqrt{1+\beta}}\mathbf{u}_j & -\sqrt{\frac{\beta}{1+\beta}}\mathbf{u}_j\end{bmatrix}}_{\triangleq\ (\mathbf{v}_{2j-1},\mathbf{v}_{2j})}\underbrace{\begin{bmatrix}\sqrt{1+\beta} & \sqrt{\frac{\beta}{1+\beta}}\\ 0 & \frac{1}{\sqrt{1+\beta}}\end{bmatrix}}_{\triangleq\ \mathbf{S}_j}.$$

Table 2: $\mathbf{B}$'s eigenpairs in descending order of eigenvalues' magnitude.

Magnitude	Eigenvalue	(generalized) Eigenvector
$\lvert\lambda_1^+\rvert = \frac{\lambda_1+\sqrt{\lambda_1^2-4\beta}}{2}$	$\lambda_1^+ \triangleq \frac{\lambda_1+\sqrt{\lambda_1^2-4\beta}}{2}$	$\mathbf{u}_1^+ \triangleq \begin{bmatrix}\lambda_1^+\mathbf{u}_1\\ \mathbf{u}_1\end{bmatrix}$
$\cdots$	$\cdots$	$\cdots$
$\lvert\lambda_s^+\rvert = \frac{\lambda_s+\sqrt{\lambda_s^2-4\beta}}{2}$	$\lambda_s^+ \triangleq \frac{\lambda_s+\sqrt{\lambda_s^2-4\beta}}{2}$	$\mathbf{u}_s^+ \triangleq \begin{bmatrix}\lambda_s^+\mathbf{u}_s\\ \mathbf{u}_s\end{bmatrix}$
$\lvert\lambda_{s+1}^\pm\rvert = \sqrt{\beta}$	$\lambda_{s+1}^\pm \triangleq \frac{\lambda_{s+1}\pm i\sqrt{4\beta-\lambda_{s+1}^2}}{2}$	$\mathbf{u}_{s+1}^\pm \triangleq \begin{bmatrix}\lambda_{s+1}^\pm\mathbf{u}_{s+1}\\ \mathbf{u}_{s+1}\end{bmatrix}$
$\cdots$	$\cdots$	$\cdots$
$\lvert\lambda_n^\pm\rvert = \sqrt{\beta}$	$\lambda_n^\pm \triangleq \frac{\lambda_n\pm i\sqrt{4\beta-\lambda_n^2}}{2}$	$\mathbf{u}_n^\pm \triangleq \begin{bmatrix}\lambda_n^\pm\mathbf{u}_n\\ \mathbf{u}_n\end{bmatrix}$
$\lvert\lambda_s^-\rvert = \frac{2\beta}{\lambda_s+\sqrt{\lambda_s^2-4\beta}}$	$\lambda_s^- \triangleq \frac{\lambda_s-\sqrt{\lambda_s^2-4\beta}}{2}$	$\mathbf{u}_s^- \triangleq \begin{bmatrix}\lambda_s^-\mathbf{u}_s\\ \mathbf{u}_s\end{bmatrix}$
$\cdots$	$\cdots$	$\cdots$
$\lvert\lambda_1^-\rvert = \frac{2\beta}{\lambda_1+\sqrt{\lambda_1^2-4\beta}}$	$\lambda_1^- \triangleq \frac{\lambda_1-\sqrt{\lambda_1^2-4\beta}}{2}$	$\mathbf{u}_1^- \triangleq \begin{bmatrix}\lambda_1^-\mathbf{u}_1\\ \mathbf{u}_1\end{bmatrix}$

Let $\mathbf{V}_\boxplus = \begin{bmatrix}\mathbf{v}_1 & \cdots & \mathbf{v}_{2n}\end{bmatrix} \in \mathbb{C}^{2n\times 2n}$ which then is unitary, i.e., $\mathbf{V}_\boxplus^{\mathrm{H}}\mathbf{V}_\boxplus = \mathbf{V}_\boxplus\mathbf{V}_\boxplus^{\mathrm{H}} = \mathbf{I}$. Thus, we can write that $\mathbf{B}\mathbf{V}_\boxplus\mathrm{diag}(\mathbf{S}_1,\cdots,\mathbf{S}_n) = \mathbf{V}_\boxplus\mathrm{diag}(\mathbf{S}_1,\cdots,\mathbf{S}_n)\widehat{\mathbf{\Lambda}}_\boxplus$, and consequently $\mathbf{B} = \mathbf{V}_\boxplus\mathrm{diag}(\mathbf{S}_1,\cdots,\mathbf{S}_n)\widehat{\mathbf{\Lambda}}_\boxplus\mathrm{diag}^{-1}(\mathbf{S}_1,\cdots,\mathbf{S}_n)\mathbf{V}_\boxplus^{\mathrm{H}}$. Let

$$\begin{aligned}\mathbf{\Lambda}_\boxplus &\triangleq \mathrm{diag}(\mathbf{S}_1,\cdots,\mathbf{S}_n)\widehat{\mathbf{\Lambda}}_\boxplus\mathrm{diag}^{-1}(\mathbf{S}_1,\cdots,\mathbf{S}_n)\\ &= \mathrm{diag}(\mathbf{S}_1\widehat{\mathbf{\Lambda}}_1\mathbf{S}_1^{-1},\cdots,\mathbf{S}_n\widehat{\mathbf{\Lambda}}_n\mathbf{S}_n^{-1}).\end{aligned}$$

If $\lambda_j \neq 2\sqrt{\beta}$, it then holds that

$$\begin{aligned}\mathbf{S}_j\widehat{\mathbf{\Lambda}}_j\mathbf{S}_j^{-1} &= \begin{bmatrix}\sqrt{1+|\lambda_j^+|^2} & \frac{1+\overline{\lambda_j^+}\lambda_j^-}{\sqrt{1+|\lambda_j^+|^2}}\\ 0 & \frac{\lambda_j^+-\lambda_j^-}{\sqrt{1+|\lambda_j^+|^2}}\end{bmatrix}\mathrm{diag}(\lambda_j^+,\lambda_j^-)\begin{bmatrix}\sqrt{1+|\lambda_j^+|^2} & \frac{1+\overline{\lambda_j^+}\lambda_j^-}{\sqrt{1+|\lambda_j^+|^2}}\\ 0 & \frac{\lambda_j^+-\lambda_j^-}{\sqrt{1+|\lambda_j^+|^2}}\end{bmatrix}^{-1}\\ &= \begin{bmatrix}\lambda_j^+ & -(1+\overline{\lambda_j^+}\lambda_j^-)\\ 0 & \lambda_j^-\end{bmatrix} = \begin{cases}\begin{bmatrix}\lambda_j^+ & -(1+\beta)\\ 0 & \lambda_j^-\end{bmatrix}, & j\le s\\ \begin{bmatrix}\lambda_j^+ & -(1+(\lambda_j^-)^2)\\ 0 & \lambda_j^-\end{bmatrix}, & \text{else}\end{cases},\end{aligned}$$

otherwise $\mathbf{S}_j\widehat{\mathbf{\Lambda}}_j\mathbf{S}_j^{-1} =$

$$\mathbf{S}_j\begin{bmatrix}\lambda_j^+ & 1\\ 0 & \lambda_j^-\end{bmatrix}\mathbf{S}_j^{-1} = \begin{bmatrix}\sqrt{1+\beta} & \sqrt{\frac{\beta}{1+\beta}}\\ 0 & \frac{1}{\sqrt{1+\beta}}\end{bmatrix}\begin{bmatrix}\sqrt{\beta} & 1\\ 0 & \sqrt{\beta}\end{bmatrix}\begin{bmatrix}\sqrt{1+\beta} & \sqrt{\frac{\beta}{1+\beta}}\\ 0 & \frac{1}{\sqrt{1+\beta}}\end{bmatrix}^{-1} = \begin{bmatrix}\sqrt{\beta} & 1+\beta\\ 0 & \sqrt{\beta}\end{bmatrix}.$$

Thus, $\mathbf{\Lambda}_{\boxplus} =$

$$\mathrm{diag}(\begin{bmatrix}\lambda_1^+ & -(1+\beta)\\ 0 & \lambda_1^-\end{bmatrix},\cdots,\begin{bmatrix}\lambda_s^+ & -(1+\beta)\\ 0 & \lambda_s^-\end{bmatrix},\begin{bmatrix}\lambda_{s+1}^+ & -(1+(\lambda_{s+1}^-)^2)\\ 0 & \lambda_{s+1}^-\end{bmatrix},\cdots,\begin{bmatrix}\lambda_n^+ & -(1+(\lambda_n^-)^2)\\ 0 & \lambda_n^-\end{bmatrix}),$$

where the j-th 2×2 upper triangular matrix for $j>s$ is replaced by $\begin{bmatrix}\sqrt{\beta} & 1+\beta\\ 0 & \sqrt{\beta}\end{bmatrix}$ if $\lambda_j = 2\sqrt{\beta}$. After certain permutations by a permutation matrix $\mathbf{P}$, $\mathbf{B}$'s Schur decomposition can be written as

$$\mathbf{B} = \mathbf{V}_{\boxplus}\mathbf{\Lambda}_{\boxplus}\mathbf{V}_{\boxplus}^{\mathrm{H}} = \mathbf{V}_{\boxplus}\mathbf{P}^\top\mathbf{P}\mathbf{\Lambda}_{\boxplus}\mathbf{P}^\top\mathbf{P}\mathbf{V}_{\boxplus}^{\mathrm{H}} = (\mathbf{V}_{\boxplus}\mathbf{P}^\top)(\mathbf{P}\mathbf{\Lambda}_{\boxplus}\mathbf{P}^\top)(\mathbf{V}_{\boxplus}\mathbf{P}^\top)^{\mathrm{H}},$$

where

$$\begin{aligned}\mathbf{V}_{\boxplus}\mathbf{P}^\top &= \begin{bmatrix}\mathbf{U}_s\mathbf{D}_s(1) & \mathbf{U}_{-s}\mathbf{D}_{-s}(1) & -\mathbf{U}_{-s}\mathbf{K}_{-s}(1) & -\mathbf{U}_s\mathbf{K}_s(1)\\ \mathbf{U}_s\mathbf{K}_s(1) & \mathbf{U}_{-s}\mathbf{K}_{-s}(1) & \mathbf{U}_{-s}\overline{\mathbf{D}_{-s}(1)} & \mathbf{U}_s\overline{\mathbf{D}_s(1)}\end{bmatrix},\\ \mathbf{P}\mathbf{\Lambda}_{\boxplus}\mathbf{P}^\top &= \begin{bmatrix}\mathbf{\Sigma}_s^+ & \mathbf{0} & \mathbf{0} & -(1+\beta)\mathbf{I}\\ \mathbf{0} & \mathbf{\Sigma}_{-s}^+ & -(\mathbf{I}+(\mathbf{\Sigma}_{-s}^-)^2) & \mathbf{0}\\ \mathbf{0} & \mathbf{0} & \mathbf{\Sigma}_{-s}^- & \mathbf{0}\\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{\Sigma}_s^-\end{bmatrix},\end{aligned}$$

$$\begin{aligned}\mathbf{D}_s(\alpha) &= \mathrm{diag}(\tfrac{\lambda_1^+}{\sqrt{\alpha^2+|\lambda_1^+|^2}},\cdots,\tfrac{\lambda_s^+}{\sqrt{\alpha^2+|\lambda_s^+|^2}}), \qquad \mathbf{D}_{-s}(\alpha) = \mathrm{diag}(\tfrac{\lambda_{s+1}^+}{\sqrt{\alpha^2+|\lambda_{s+1}^+|^2}},\cdots,\tfrac{\lambda_n^+}{\sqrt{\alpha^2+|\lambda_n^+|^2}}),\\ \mathbf{K}_s(\alpha) &= \mathrm{diag}(\tfrac{\alpha}{\sqrt{\alpha^2+|\lambda_1^+|^2}},\cdots,\tfrac{\alpha}{\sqrt{\alpha^2+|\lambda_s^+|^2}}), \qquad \mathbf{K}_{-s}(\alpha) = \mathrm{diag}(\tfrac{\alpha}{\sqrt{\alpha^2+|\lambda_{s+1}^+|^2}},\cdots,\tfrac{\alpha}{\sqrt{\alpha^2+|\lambda_n^+|^2}}),\end{aligned}$$

$$\begin{aligned}\mathbf{\Sigma}_s^+ &= \mathrm{diag}(\lambda_1^+,\cdots,\lambda_s^+), \qquad \mathbf{\Sigma}_{-s}^+ = \mathrm{diag}(\lambda_{s+1}^+,\cdots,\lambda_n^+),\\ \mathbf{\Sigma}_s^- &= \mathrm{diag}(\lambda_1^-,\cdots,\lambda_s^-), \qquad \mathbf{\Sigma}_{-s}^- = \mathrm{diag}(\lambda_{s+1}^-,\cdots,\lambda_n^-).\end{aligned}$$

□

Lemma 2.2 If there exists an integer s such that $\lambda_s > 2\sqrt{\beta} \geq \lambda_{s+1}$ and $\lambda_n > 0$, we then have the following Schur decomposition for $\mathbf{B}^{\mathrm{H}} = \begin{bmatrix}\mathbf{A} & \mathbf{I}\\ -\beta\mathbf{I} & \mathbf{0}\end{bmatrix}$ with notations given in Lemma 2.1:

$$\begin{bmatrix}\mathbf{A} & \mathbf{I}\\ -\beta\mathbf{I} & \mathbf{0}\end{bmatrix} = \begin{bmatrix}\mathbf{U}_s\mathbf{D}_s(\beta) & \mathbf{U}_{-s}\mathbf{D}_{-s}(\beta) & \mathbf{U}_{-s}\mathbf{K}_{-s}(\beta) & \mathbf{U}_s\mathbf{K}_s(\beta)\\ -\mathbf{U}_s\mathbf{K}_s(\beta) & -\mathbf{U}_{-s}\mathbf{K}_{-s}(\beta) & \mathbf{U}_{-s}\overline{\mathbf{D}_{-s}(\beta)} & \mathbf{U}_s\overline{\mathbf{D}_s(\beta)}\end{bmatrix} \times$$
$$\begin{bmatrix}\mathbf{\Sigma}_s^+ & \mathbf{0} & \mathbf{0} & (1+\beta)\mathbf{I}\\ \mathbf{0} & \mathbf{\Sigma}_{-s}^+ & \mathbf{T}(\beta) & \mathbf{0}\\ \mathbf{0} & \mathbf{0} & \mathbf{\Sigma}_{-s}^- & \mathbf{0}\\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{\Sigma}_s^-\end{bmatrix}\begin{bmatrix}\mathbf{U}_s\mathbf{D}_s(\beta) & \mathbf{U}_{-s}\mathbf{D}_{-s}(\beta) & \mathbf{U}_{-s}\mathbf{K}_{-s}(\beta) & \mathbf{U}_s\mathbf{K}_s(\beta)\\ -\mathbf{U}_s\mathbf{K}_s(\beta) & -\mathbf{U}_{-s}\mathbf{K}_{-s}(\beta) & \mathbf{U}_{-s}\overline{\mathbf{D}_{-s}(\beta)} & \mathbf{U}_s\overline{\mathbf{D}_s(\beta)}\end{bmatrix}^{\mathrm{H}},$$

where, if there exists any $j>s$ such that $\lambda_j = 2\sqrt{\beta}$, then $\lambda_j^{\pm} = \sqrt{\beta}$ and the corresponding diagonal entry in the block $\mathbf{T}(\beta)$ is replaced with $(1+\beta)$.

Proof The proof is almost the same as the above one. We only present the differences. First of all, note that $\mathbf{B}$ and $\mathbf{B}^{\mathrm{H}}$ have exactly the same spectrum. Suppose one eigenpair of $\mathbf{A}$ is $(\lambda_j, \mathbf{u}_j)$. We then have $\mathbf{B}^{\mathrm{H}}$'s eigenvalue $\lambda_j^{\pm}$ which are the roots of the quadratic equation $\mu^2 - \lambda_j \mu + \beta = 0$ in μ. We then have that $\begin{bmatrix} \mathbf{A} & \mathbf{I} \\ -\beta\mathbf{I} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mu\mathbf{u}_j \\ -\beta\mathbf{u}_j \end{bmatrix} = \mu \begin{bmatrix} \mu\mathbf{u}_j \\ -\beta\mathbf{u}_j \end{bmatrix}$. Thus, a double eigenpair of $\mathbf{B}^{\mathrm{H}}$ then can be written as $\left(\frac{\lambda_j \pm \sqrt{\lambda_j^2 - 4\beta}}{2}, \begin{bmatrix} \frac{\lambda_j \pm \sqrt{\lambda_j^2 - 4\beta}}{2} \mathbf{u}_j \\ -\beta\mathbf{u}_j \end{bmatrix} \right)$. For $\lambda_s > 2\sqrt{\beta} \geq \lambda_{s+1}$, all the eigenvectors of $\mathbf{B}^{\mathrm{H}}$ in the above double eigenpairs together (i.e., $j = 1, \cdots, n$) will span the whole Euclidean complex vector space $\mathbb{C}^{2n}$ as long as there exists no $j > s$ such that $\lambda_j = 2\sqrt{\beta}$. In the collapsed case, for each j of the kind, we additionally derive one generalized eigenvector $\mathbf{z}_j \in \mathbb{R}^{2n \times 1}$ of $\mathbf{B}^{\mathrm{H}}$ from the collapsed single eigenpair to span $\mathbb{C}^{2n}$, i.e., $(\mathbf{B}^{\mathrm{H}} - \sqrt{\beta}\mathbf{I})\mathbf{z}_j = \begin{bmatrix} \sqrt{\beta}\mathbf{u}_j \\ -\beta\mathbf{u}_j \end{bmatrix}$. Again, $\mathbf{z}_j = \begin{bmatrix} \mathbf{u}_j \\ \mathbf{0} \end{bmatrix}$. $\mathbf{B}^{\mathrm{H}}$'s eigenvalues as well as corresponding eigenvectors or generalized eigenvectors in descending order of eigenvalues' magnitude are listed in Table 3 where, if there is any $j > s$ such that $\lambda_j = 2\sqrt{\beta}$ then the corresponding row in the middle is replaced by the following one:

$$|\lambda_j^{\pm}| = \sqrt{\beta}, \quad \lambda_j^{\pm} \triangleq \sqrt{\beta}, \quad \mathbf{u}_j^{+} \triangleq \begin{bmatrix} \sqrt{\beta}\mathbf{u}_j \\ -\beta\mathbf{u}_j \end{bmatrix}, \quad \mathbf{u}_j^{-} \triangleq \begin{bmatrix} \mathbf{u}_j \\ \mathbf{0} \end{bmatrix}.$$

Thus, we have $\mathbf{B}^{\mathrm{H}}\mathbf{u}_j^{\pm} = \lambda_j^{\pm}\mathbf{u}_j^{\pm}$ for j such that $\lambda_j \neq 2\sqrt{\beta}$, otherwise $\begin{cases} \mathbf{B}^{\mathrm{H}}\mathbf{u}_j^{+} = \lambda_j^{+}\mathbf{u}_j^{+}, \\ \mathbf{B}^{\mathrm{H}}\mathbf{u}_j^{-} = \mathbf{u}_j^{+} + \lambda_j^{-}\mathbf{u}_j^{-}. \end{cases}$

Table 3: $\mathbf{B}^{\mathrm{H}}$'s eigenpairs in descending order of eigenvalues' magnitude.

Magnitude	Eigenvalue	(generalized) Eigenvector
$\lvert\lambda_1^{+}\rvert = \frac{\lambda_1 + \sqrt{\lambda_1^2 - 4\beta}}{2}$	$\lambda_1^{+} \triangleq \frac{\lambda_1 + \sqrt{\lambda_1^2 - 4\beta}}{2}$	$\mathbf{u}_1^{+} \triangleq \begin{bmatrix} \lambda_1^{+}\mathbf{u}_1 \\ -\beta\mathbf{u}_1 \end{bmatrix}$
$\cdots$	$\cdots$	$\cdots$
$\lvert\lambda_s^{+}\rvert = \frac{\lambda_s + \sqrt{\lambda_s^2 - 4\beta}}{2}$	$\lambda_s^{+} \triangleq \frac{\lambda_s + \sqrt{\lambda_s^2 - 4\beta}}{2}$	$\mathbf{u}_s^{+} \triangleq \begin{bmatrix} \lambda_s^{+}\mathbf{u}_s \\ -\beta\mathbf{u}_s \end{bmatrix}$
$\lvert\lambda_{s+1}^{\pm}\rvert = \sqrt{\beta}$	$\lambda_{s+1}^{\pm} \triangleq \frac{\lambda_{s+1} \pm i\sqrt{4\beta - \lambda_{s+1}^2}}{2}$	$\mathbf{u}_{s+1}^{\pm} \triangleq \begin{bmatrix} \lambda_{s+1}^{\pm}\mathbf{u}_{s+1} \\ -\beta\mathbf{u}_{s+1} \end{bmatrix}$
$\cdots$	$\cdots$	$\cdots$
$\lvert\lambda_n^{\pm}\rvert = \sqrt{\beta}$	$\lambda_n^{\pm} \triangleq \frac{\lambda_n \pm i\sqrt{4\beta - \lambda_n^2}}{2}$	$\mathbf{u}_n^{\pm} \triangleq \begin{bmatrix} \lambda_n^{\pm}\mathbf{u}_n \\ -\beta\mathbf{u}_n \end{bmatrix}$
$\lvert\lambda_s^{-}\rvert = \frac{2\beta}{\lambda_s + \sqrt{\lambda_s^2 - 4\beta}}$	$\lambda_s^{-} \triangleq \frac{\lambda_s - \sqrt{\lambda_s^2 - 4\beta}}{2}$	$\mathbf{u}_s^{-} \triangleq \begin{bmatrix} \lambda_s^{-}\mathbf{u}_s \\ -\beta\mathbf{u}_s \end{bmatrix}$
$\cdots$	$\cdots$	$\cdots$
$\lvert\lambda_1^{-}\rvert = \frac{2\beta}{\lambda_1 + \sqrt{\lambda_1^2 - 4\beta}}$	$\lambda_1^{-} \triangleq \frac{\lambda_1 - \sqrt{\lambda_1^2 - 4\beta}}{2}$	$\mathbf{u}_1^{-} \triangleq \begin{bmatrix} \lambda_1^{-}\mathbf{u}_1 \\ -\beta\mathbf{u}_1 \end{bmatrix}$

In matrix form, $\mathbf{B}^{\mathrm{H}}\widehat{\mathbf{V}}_{\boxplus} = \widehat{\mathbf{V}}_{\boxplus}\widehat{\mathbf{\Lambda}}_{\boxplus}$, where

$$\widehat{\mathbf{V}}_{\boxplus} = \begin{bmatrix}\mathbf{u}_1^+ & \mathbf{u}_1^- & \cdots & \mathbf{u}_n^+ & \mathbf{u}_n^-\end{bmatrix} \quad , \quad \widehat{\mathbf{\Lambda}}_{\boxplus} = \mathrm{diag}(\widehat{\mathbf{\Lambda}}_1, \cdots, \widehat{\mathbf{\Lambda}}_n) \quad , \quad \widehat{\mathbf{\Lambda}}_j = \mathrm{diag}(\lambda_j^+, \lambda_j^-),$$

where for all j such that $\lambda_j = 2\sqrt{\beta}$, each of the corresponding 2×2 diagonal blocks is replaced with the 2×2 upper triangular block $\widehat{\mathbf{\Lambda}}_j = \begin{bmatrix}\lambda_j^+ & 1 \\ 0 & \lambda_j^-\end{bmatrix}$. We now orthonormalize the pair $(\mathbf{u}_j^+, \mathbf{u}_j^-)$ for each $j = 1, \cdots, n$. If $\lambda_j \neq 2\sqrt{\beta}$, then we have the orthonormalization:

$$\begin{bmatrix}\mathbf{u}_j^+ & \mathbf{u}_j^-\end{bmatrix} = \begin{bmatrix}\lambda_j^+\mathbf{u}_j & \lambda_j^-\mathbf{u}_j \\ -\beta\mathbf{u}_j & -\beta\mathbf{u}_j\end{bmatrix} = \underbrace{\begin{bmatrix} \frac{\lambda_j^+}{\sqrt{\beta^2+|\lambda_j^+|^2}}\mathbf{u}_j & \frac{\beta}{\sqrt{\beta^2+|\lambda_j^+|^2}}\mathbf{u}_j \\ -\frac{\beta}{\sqrt{\beta^2+|\lambda_j^+|^2}}\mathbf{u}_j & \frac{\overline{\lambda_j^+}}{\sqrt{\beta^2+|\lambda_j^+|^2}}\mathbf{u}_j \end{bmatrix}}_{\triangleq\ (\mathbf{v}_{2j-1}, \mathbf{v}_{2j})} \underbrace{\begin{bmatrix} \sqrt{\beta^2+|\lambda_j^+|^2} & \frac{\beta^2+\overline{\lambda_j^+}\lambda_j^-}{\sqrt{\beta^2+|\lambda_j^+|^2}} \\ 0 & \frac{(\lambda_j^- - \lambda_j^+)\beta}{\sqrt{\beta^2+|\lambda_j^+|^2}} \end{bmatrix}}_{\triangleq\ \mathbf{S}_j}.$$

If $\lambda_j = 2\sqrt{\beta}$ then the orthonormalization is

$$\begin{bmatrix}\mathbf{u}_j^+ & \mathbf{u}_j^-\end{bmatrix} = \begin{bmatrix}\sqrt{\beta}\mathbf{u}_j & \mathbf{u}_j \\ -\beta\mathbf{u}_j & \mathbf{0}\end{bmatrix} = \underbrace{\begin{bmatrix} \frac{1}{\sqrt{1+\beta}}\mathbf{u}_j & \sqrt{\frac{\beta}{1+\beta}}\mathbf{u}_j \\ -\sqrt{\frac{\beta}{1+\beta}}\mathbf{u}_j & \frac{1}{\sqrt{1+\beta}}\mathbf{u}_j \end{bmatrix}}_{\triangleq\ (\mathbf{v}_{2j-1}, \mathbf{v}_{2j})} \underbrace{\begin{bmatrix} \sqrt{\beta(1+\beta)} & \frac{1}{\sqrt{1+\beta}} \\ 0 & \sqrt{\frac{\beta}{1+\beta}} \end{bmatrix}}_{\triangleq\ \mathbf{S}_j}.$$

Let $\mathbf{V}_{\boxplus} = \begin{bmatrix}\mathbf{v}_1 & \cdots & \mathbf{v}_{2n}\end{bmatrix} \in \mathbb{C}^{2n \times 2n}$ which then is unitary, i.e., $\mathbf{V}_{\boxplus}^{\mathrm{H}}\mathbf{V}_{\boxplus} = \mathbf{V}_{\boxplus}\mathbf{V}_{\boxplus}^{\mathrm{H}} = \mathbf{I}$. Thus, we can write that $\mathbf{B}^{\mathrm{H}}\mathbf{V}_{\boxplus}\mathrm{diag}(\mathbf{S}_1, \cdots, \mathbf{S}_n) = \mathbf{V}_{\boxplus}\mathrm{diag}(\mathbf{S}_1, \cdots, \mathbf{S}_n)\widehat{\mathbf{\Lambda}}_{\boxplus}$, and consequently $\mathbf{B}^{\mathrm{H}} = \mathbf{V}_{\boxplus}\mathrm{diag}(\mathbf{S}_1, \cdots, \mathbf{S}_n)\widehat{\mathbf{\Lambda}}_{\boxplus}\mathrm{diag}^{-1}(\mathbf{S}_1, \cdots, \mathbf{S}_n)\mathbf{V}_{\boxplus}^{\mathrm{H}}$. Let

$$\mathbf{\Lambda}_{\boxplus} \triangleq \mathrm{diag}(\mathbf{S}_1, \cdots, \mathbf{S}_n)\widehat{\mathbf{\Lambda}}_{\boxplus}\mathrm{diag}^{-1}(\mathbf{S}_1, \cdots, \mathbf{S}_n) = \mathrm{diag}(\mathbf{S}_1\widehat{\mathbf{\Lambda}}_1\mathbf{S}_1^{-1}, \cdots, \mathbf{S}_n\widehat{\mathbf{\Lambda}}_n\mathbf{S}_n^{-1}).$$

If $\lambda_j \neq 2\sqrt{\beta}$, it then holds that

$$\begin{aligned}\mathbf{S}_j\widehat{\mathbf{\Lambda}}_j\mathbf{S}_j^{-1} &= \begin{bmatrix} \sqrt{\beta^2+|\lambda_j^+|^2} & \frac{\beta^2+\overline{\lambda_j^+}\lambda_j^-}{\sqrt{\beta^2+|\lambda_j^+|^2}} \\ 0 & \frac{(\lambda_j^- - \lambda_j^+)\beta}{\sqrt{\beta^2+|\lambda_j^+|^2}} \end{bmatrix} \mathrm{diag}(\lambda_j^+, \lambda_j^-) \begin{bmatrix} \sqrt{\beta^2+|\lambda_j^+|^2} & \frac{\beta^2+\overline{\lambda_j^+}\lambda_j^-}{\sqrt{\beta^2+|\lambda_j^+|^2}} \\ 0 & \frac{(\lambda_j^- - \lambda_j^+)\beta}{\sqrt{\beta^2+|\lambda_j^+|^2}} \end{bmatrix}^{-1} \\ &= \begin{bmatrix}\lambda_j^+ & \beta + \frac{1}{\beta}\overline{\lambda_j^+}\lambda_j^- \\ 0 & \lambda_j^-\end{bmatrix} = \begin{cases} \begin{bmatrix}\lambda_j^+ & 1+\beta \\ 0 & \lambda_j^-\end{bmatrix}, & j \leq s \\ \begin{bmatrix}\lambda_j^+ & \beta + \frac{1}{\beta}(\lambda_j^-)^2 \\ 0 & \lambda_j^-\end{bmatrix}, & \text{else} \end{cases},\end{aligned}$$

otherwise $\mathbf{S}_j\widehat{\mathbf{\Lambda}}_j\mathbf{S}_j^{-1} =$

$$\mathbf{S}_j\begin{bmatrix}\lambda_j^+ & 1 \\ 0 & \lambda_j^-\end{bmatrix}\mathbf{S}_j^{-1} = \begin{bmatrix} \sqrt{\beta(1+\beta)} & \frac{1}{\sqrt{1+\beta}} \\ 0 & \sqrt{\frac{\beta}{1+\beta}} \end{bmatrix} \begin{bmatrix}\sqrt{\beta} & 1 \\ 0 & \sqrt{\beta}\end{bmatrix} \begin{bmatrix} \sqrt{\beta(1+\beta)} & \frac{1}{\sqrt{1+\beta}} \\ 0 & \sqrt{\frac{\beta}{1+\beta}} \end{bmatrix}^{-1} = \begin{bmatrix}\sqrt{\beta} & 1+\beta \\ 0 & \sqrt{\beta}\end{bmatrix}.$$

Thus,

$$\mathbf{\Lambda}_{\boxplus} = \mathrm{diag}(\begin{bmatrix}\lambda_1^+ & 1+\beta \\ 0 & \lambda_1^-\end{bmatrix}, \cdots, \begin{bmatrix}\lambda_s^+ & 1+\beta \\ 0 & \lambda_s^-\end{bmatrix}, \begin{bmatrix}\lambda_{s+1}^+ & \beta+\frac{1}{\beta}(\lambda_{s+1}^-)^2 \\ 0 & \lambda_{s+1}^-\end{bmatrix}, \cdots, \begin{bmatrix}\lambda_n^+ & \beta+\frac{1}{\beta}(\lambda_n^-)^2 \\ 0 & \lambda_n^-\end{bmatrix}),$$

where the j-th 2×2 upper triangular matrix for $j > s$ is replaced by $\begin{bmatrix}\sqrt{\beta} & 1+\beta \\ 0 & \sqrt{\beta}\end{bmatrix}$ if $\lambda_j = 2\sqrt{\beta}$.

Therefore, $\mathbf{B}^{\mathrm{H}}$'s Schur decomposition can be written as

$$\mathbf{B}^{\mathrm{H}} = \mathbf{V}_{\boxplus}\mathbf{\Lambda}_{\boxplus}\mathbf{V}_{\boxplus}^{\mathrm{H}} = \mathbf{V}_{\boxplus}\mathbf{P}^\top\mathbf{P}\mathbf{\Lambda}_{\boxplus}\mathbf{P}^\top\mathbf{P}\mathbf{V}_{\boxplus}^{\mathrm{H}} = (\mathbf{V}_{\boxplus}\mathbf{P}^\top)(\mathbf{P}\mathbf{\Lambda}_{\boxplus}\mathbf{P}^\top)(\mathbf{V}_{\boxplus}\mathbf{P}^\top)^{\mathrm{H}},$$

where

$$\begin{aligned}
\mathbf{V}_{\boxplus}\mathbf{P}^\top &= \begin{bmatrix}\mathbf{U}_s\mathbf{D}_s(\beta) & \mathbf{U}_{-s}\mathbf{D}_{-s}(\beta) & \mathbf{U}_{-s}\mathbf{K}_{-s}(\beta) & \mathbf{U}_s\mathbf{K}_s(\beta) \\ -\mathbf{U}_s\mathbf{K}_s(\beta) & -\mathbf{U}_{-s}\mathbf{K}_{-s}(\beta) & \mathbf{U}_{-s}\overline{\mathbf{D}_{-s}(\beta)} & \mathbf{U}_s\overline{\mathbf{D}_s(\beta)}\end{bmatrix}, \\
\mathbf{P}\mathbf{\Lambda}_{\boxplus}\mathbf{P}^\top &= \begin{bmatrix}\mathbf{\Sigma}_s^+ & \mathbf{0} & \mathbf{0} & (1+\beta)\mathbf{I} \\ \mathbf{0} & \mathbf{\Sigma}_{-s}^+ & \mathbf{T}(\beta) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{\Sigma}_{-s}^- & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{\Sigma}_s^-\end{bmatrix}.
\end{aligned}$$

□

Lemma 3 $\mathbf{\Omega} = -(1+\beta)\begin{bmatrix}\mathbf{0}_{q\times(n-q)} & \mathbf{0}_{q\times(n-q)} & \mathrm{diag}(\frac{1}{\sqrt{\lambda_1^2-4\beta}}, \cdots, \frac{1}{\sqrt{\lambda_q^2-4\beta}})\end{bmatrix}$ and $\|\mathbf{\Omega}\|_2 = \frac{1+\beta}{\sqrt{\lambda_q^2-4\beta}}$.

Proof Since $\mathbf{\Lambda}_q$ and $\mathbf{\Lambda}_{-q}$ have their spectra divided by $\sqrt{\beta}$ in magnitude, by Theorem VII.2.2 in Bhatia (1997), the equation $\mathbf{\Lambda}_q\mathbf{\Omega} - \mathbf{\Omega}\mathbf{\Lambda}_{-q} = -\mathbf{\Lambda}_{q,-q}$ has the solution $\mathbf{\Omega} = -\sum_{j=0}^{\infty}\mathbf{\Lambda}_q^{-j-1}\mathbf{\Lambda}_{q,-q}\mathbf{\Lambda}_{-q}^j$, which can be simplified as follows. Noting Lemma 2 and Eq. (5) about expressions of $\mathbf{\Lambda}_q$, $\mathbf{\Lambda}_{-q}$, and $\mathbf{\Lambda}_{q,-q}$, we have that

$$\begin{aligned}
\mathbf{\Omega} &= -\sum_{j=0}^{\infty}\mathbf{\Lambda}_q^{-j-1}\mathbf{\Lambda}_{q,-q}\mathbf{\Lambda}_{-q}^j = -(1+\beta)\sum_{j=0}^{\infty}\mathbf{\Lambda}_q^{-j-1}(\begin{bmatrix}\mathbf{0} & \mathbf{0} & \mathbf{I}\end{bmatrix}\mathbf{\Lambda}_{-q}^j) \\
&= -(1+\beta)\sum_{j=0}^{\infty}\begin{bmatrix}\mathbf{0} & \mathbf{0} & (\mathbf{\Sigma}_q^+)^{-j-1}(\mathbf{\Sigma}_q^-)^j\end{bmatrix} = -(1+\beta)(\mathbf{\Sigma}_q^+)^{-1}\sum_{j=0}^{\infty}\begin{bmatrix}\mathbf{0} & \mathbf{0} & ((\mathbf{\Sigma}_q^+)^{-1}(\mathbf{\Sigma}_q^-))^j\end{bmatrix} \\
&= -(1+\beta)\begin{bmatrix}\mathbf{0} & \mathbf{0} & \mathrm{diag}(\frac{1}{\lambda_1^+}\sum_{j=0}^{\infty}(\frac{\lambda_1^-}{\lambda_1^+})^j, \cdots, \frac{1}{\lambda_q^+}\sum_{j=0}^{\infty}(\frac{\lambda_q^-}{\lambda_q^+})^j)\end{bmatrix} \\
&= -(1+\beta)\begin{bmatrix}\mathbf{0} & \mathbf{0} & \mathrm{diag}(\frac{\frac{1}{\lambda_1^+}}{1-\frac{\lambda_1^-}{\lambda_1^+}}, \cdots, \frac{\frac{1}{\lambda_q^+}}{1-\frac{\lambda_q^-}{\lambda_q^+}})\end{bmatrix} = -(1+\beta)\begin{bmatrix}\mathbf{0} & \mathbf{0} & \mathrm{diag}(\frac{1}{\lambda_1^+-\lambda_1^-}, \cdots, \frac{1}{\lambda_q^+-\lambda_q^-})\end{bmatrix} \\
&= -(1+\beta)\begin{bmatrix}\mathbf{0} & \mathbf{0} & \mathrm{diag}(\frac{1}{\sqrt{\lambda_1^2-4\beta}}, \cdots, \frac{1}{\sqrt{\lambda_q^2-4\beta}})\end{bmatrix}.
\end{aligned}$$

It is easy to see that $\|\mathbf{\Omega}\|_2 = \frac{1+\beta}{\sqrt{\lambda_q^2-4\beta}}$. □

Lemma 4 Under the noise conditions $\|\boldsymbol{\xi}_t\|_2 \leq \frac{1}{256}\frac{(\lambda_k^+-\sqrt{\beta})^2}{(1+l_1)(1+\beta)}\sin\theta_{\max}(\mathbf{Y}_t,\mathbf{V}_q)$, $\|\mathbf{U}_q^\top\boldsymbol{\xi}_t\|_2 \leq \frac{1}{16}\frac{\lambda_q^+(\lambda_k^+-\sqrt{\beta})}{(1+l_2)(1+\lambda_k^+)}\cos\theta_{\max}(\mathbf{Y}_t,\mathbf{W}_q)$, and

$$\mathrm{col}((\mathbf{W}_q^{\mathrm{H}}\mathbf{Y}_t+\mathrm{diag}^{-1}(\lambda_1^+,\cdots,\lambda_q^+)\mathbf{W}_q^{\mathrm{H}}\mathbf{E}_t)^\dagger\begin{bmatrix}\mathbf{I}_k\\ \mathbf{0}\end{bmatrix})\subset\mathrm{col}((\mathbf{W}_q^{\mathrm{H}}\mathbf{Y}_t)^\dagger\begin{bmatrix}\mathbf{I}_k\\ \mathbf{0}\end{bmatrix}),$$

it holds that

$$\Phi_{t+1}\leq(1-\frac{1}{4}\sqrt{\frac{\lambda_k-2\sqrt{\beta}}{\lambda_k}})\Phi_t.$$

Proof Let $\mathbf{H}_t=(\boldsymbol{\Gamma}\mathbf{Q}\mathbf{V}_{-q}^{\mathrm{H}}\mathbf{Y}_t)(\mathbf{W}_q^{\mathrm{H}}\mathbf{Y}_t)^\dagger\begin{bmatrix}\mathbf{I}_k\\ \mathbf{0}\end{bmatrix}$. We start from expanding $\mathbf{H}_{t+1}$ in a way different from Balcan et al. (2016) to analyze $\Phi_{t+1}=\|\mathbf{H}_{t+1}\|_2$. By Eq. (10), we have that

$$\begin{aligned}\mathbf{H}_{t+1} &= (\boldsymbol{\Gamma}\mathbf{Q}\mathbf{V}_{-q}^{\mathrm{H}}\mathbf{Y}_{t+1})(\mathbf{W}_q^{\mathrm{H}}\mathbf{Y}_{t+1})^\dagger\begin{bmatrix}\mathbf{I}_k\\ \mathbf{0}\end{bmatrix}\\ &= (\boldsymbol{\Gamma}\mathbf{Q}\boldsymbol{\Lambda}_{-q}\mathbf{V}_{-q}^{\mathrm{H}}\mathbf{Y}_t+\boldsymbol{\Gamma}\mathbf{Q}\mathbf{V}_{-q}^{\mathrm{H}}\mathbf{E}_t)(\boldsymbol{\Lambda}_q\mathbf{W}_q^{\mathrm{H}}\mathbf{Y}_t+\mathbf{W}_q^{\mathrm{H}}\mathbf{E}_t)^\dagger\begin{bmatrix}\mathbf{I}_k\\ \mathbf{0}\end{bmatrix}.\end{aligned}$$

For brevity, let

$$\mathbf{Z}_1=\boldsymbol{\Gamma}\mathbf{Q}\mathbf{V}_{-q}^{\mathrm{H}}\mathbf{Y}_t,\quad \mathbf{Z}_2=\mathbf{W}_q^{\mathrm{H}}\mathbf{Y}_t.$$

Then $\mathbf{H}_t=\mathbf{Z}_1\mathbf{Z}_2^\dagger\begin{bmatrix}\mathbf{I}_k\\ \mathbf{0}\end{bmatrix}$. For the above equation about $\mathbf{H}_{t+1}$, on one hand,

$$\boldsymbol{\Gamma}\mathbf{Q}\boldsymbol{\Lambda}_{-q}\mathbf{V}_{-q}^{\mathrm{H}}\mathbf{Y}_t+\boldsymbol{\Gamma}\mathbf{Q}\mathbf{V}_{-q}^{\mathrm{H}}\mathbf{E}_t=(\boldsymbol{\Gamma}\mathbf{Q}\boldsymbol{\Lambda}_{-q}(\boldsymbol{\Gamma}\mathbf{Q})^{-1}+\boldsymbol{\Gamma}\mathbf{Q}\mathbf{V}_{-q}^{\mathrm{H}}\mathbf{E}_t(\mathbf{V}_{-q}^{\mathrm{H}}\mathbf{Y}_t)^\dagger(\boldsymbol{\Gamma}\mathbf{Q})^{-1})\mathbf{Z}_1,$$

where $(\mathbf{V}_{-q}^{\mathrm{H}}\mathbf{Y}_t)^\dagger(\mathbf{V}_{-q}^{\mathrm{H}}\mathbf{Y}_t)=\mathbf{I}_p$. On the other hand,

$$(\boldsymbol{\Lambda}_q\mathbf{W}_q^{\mathrm{H}}\mathbf{Y}_t+\mathbf{W}_q^{\mathrm{H}}\mathbf{E}_t)^\dagger\begin{bmatrix}\mathbf{I}_k\\ \mathbf{0}\end{bmatrix}=(\mathbf{Z}_2+\boldsymbol{\Lambda}_q^{-1}\mathbf{W}_q^{\mathrm{H}}\mathbf{E}_t)^\dagger\boldsymbol{\Lambda}_q^{-1}\begin{bmatrix}\mathbf{I}_k\\ \mathbf{0}\end{bmatrix}=(\mathbf{Z}_2+\boldsymbol{\Lambda}_q^{-1}\mathbf{W}_q^{\mathrm{H}}\mathbf{E}_t)^\dagger\begin{bmatrix}\mathbf{I}_k\\ \mathbf{0}\end{bmatrix}\boldsymbol{\Lambda}_k^{-1}$$

where $\boldsymbol{\Lambda}_k$ is the $k\times k$ leading principal submatrix of $\boldsymbol{\Lambda}_q$. To proceed, let

$$(\mathbf{Z}_2+\boldsymbol{\Lambda}_q^{-1}\mathbf{W}_q^{\mathrm{H}}\mathbf{E}_t)^\dagger\begin{bmatrix}\mathbf{I}_k\\ \mathbf{0}\end{bmatrix}=\mathbf{Z}_2^\dagger\begin{bmatrix}\mathbf{I}_k\\ \mathbf{0}\end{bmatrix}\boldsymbol{\Psi}.$$

By the last noise assumption, the above equation has a solution for $\boldsymbol{\Psi}\in\mathbb{R}^{k\times k}$. Pre-multiplying both sides of the above equation by $\begin{bmatrix}\mathbf{I}_k & \mathbf{0}\end{bmatrix}\mathbf{Z}_2$, we get that

$$\boldsymbol{\Psi}=\begin{bmatrix}\mathbf{I}_k & \mathbf{0}\end{bmatrix}\mathbf{Z}_2(\mathbf{Z}_2+\boldsymbol{\Lambda}_q^{-1}\mathbf{W}_q^{\mathrm{H}}\mathbf{E}_t)^\dagger\begin{bmatrix}\mathbf{I}_k\\ \mathbf{0}\end{bmatrix},$$

where we have used that $\mathbf{Z}_2\mathbf{Z}_2^\dagger=\mathbf{I}_q$. We thus can write that

$$\begin{aligned}\mathbf{H}_{t+1} &= (\boldsymbol{\Gamma}\mathbf{Q}\boldsymbol{\Lambda}_{-q}\mathbf{V}_{-q}^{\mathrm{H}}\mathbf{Y}_t+\boldsymbol{\Gamma}\mathbf{Q}\mathbf{V}_{-q}^{\mathrm{H}}\mathbf{E}_t)(\boldsymbol{\Lambda}_q\mathbf{W}_q^{\mathrm{H}}\mathbf{Y}_t+\mathbf{W}_q^{\mathrm{H}}\mathbf{E}_t)^\dagger\begin{bmatrix}\mathbf{I}_k\\ \mathbf{0}\end{bmatrix}\\ &= (\boldsymbol{\Gamma}\mathbf{Q}\boldsymbol{\Lambda}_{-q}(\boldsymbol{\Gamma}\mathbf{Q})^{-1}+\boldsymbol{\Gamma}\mathbf{Q}\mathbf{V}_{-q}^{\mathrm{H}}\mathbf{E}_t(\mathbf{V}_{-q}^{\mathrm{H}}\mathbf{Y}_t)^\dagger(\boldsymbol{\Gamma}\mathbf{Q})^{-1})\mathbf{Z}_1\mathbf{Z}_2^\dagger\begin{bmatrix}\mathbf{I}_k\\ \mathbf{0}\end{bmatrix}\cdot\\ &\quad\begin{bmatrix}\mathbf{I}_k & \mathbf{0}\end{bmatrix}\mathbf{Z}_2(\mathbf{Z}_2+\boldsymbol{\Lambda}_q^{-1}\mathbf{W}_q^{\mathrm{H}}\mathbf{E}_t)^\dagger\begin{bmatrix}\mathbf{I}_k\\ \mathbf{0}\end{bmatrix}\boldsymbol{\Lambda}_k^{-1}\end{aligned}$$

$$= (\mathbf{\Gamma Q \Lambda}_{-q}(\mathbf{\Gamma Q})^{-1} + \mathbf{\Gamma Q V}_{-q}^{\mathrm{H}}\mathbf{E}_t(\mathbf{V}_{-q}^{\mathrm{H}}\mathbf{Y}_t)^{\dagger}(\mathbf{\Gamma Q})^{-1})\mathbf{H}_t$$
$$\cdot \begin{bmatrix}\mathbf{I}_k & \mathbf{0}\end{bmatrix}\mathbf{Z}_2(\mathbf{Z}_2 + \mathbf{\Lambda}_q^{-1}\mathbf{W}_q^{\mathrm{H}}\mathbf{E}_t)^{\dagger}\begin{bmatrix}\mathbf{I}_k \\ \mathbf{0}\end{bmatrix}\mathbf{\Lambda}_k^{-1}.$$

By the definition of the pseudo inverse of a matrix, we have that

$$\|\mathbf{H}_{t+1}\|_2 \le \|\mathbf{H}_t\|_2\|\mathbf{\Lambda}_k^{-1}\|_2 \underbrace{\|\mathbf{\Gamma Q \Lambda}_{-q}(\mathbf{\Gamma Q})^{-1} + \mathbf{\Gamma Q V}_{-q}^{\mathrm{H}}\mathbf{E}_t(\mathbf{V}_{-q}^{\mathrm{H}}\mathbf{Y}_t)^{\dagger}(\mathbf{\Gamma Q})^{-1}\|_2}_{\triangleq\ \varphi_1} \times$$
$$\underbrace{\|\begin{bmatrix}\mathbf{I}_k & \mathbf{0}\end{bmatrix}\mathbf{Z}_2(\mathbf{Z}_2 + \mathbf{\Lambda}_q^{-1}\mathbf{W}_q^{\mathrm{H}}\mathbf{E}_t)^{\top}((\mathbf{Z}_2 + \mathbf{\Lambda}_q^{-1}\mathbf{W}_q^{\mathrm{H}}\mathbf{E}_t)(\mathbf{Z}_2 + \mathbf{\Lambda}_q^{-1}\mathbf{W}_q^{\mathrm{H}}\mathbf{E}_t)^{\top})^{-1}\begin{bmatrix}\mathbf{I}_k \\ \mathbf{0}\end{bmatrix}\|_2}_{\triangleq\ \varphi_2}.$$

Noting that $\|(\mathbf{V}_{-q}^{\mathrm{H}}\mathbf{Y}_t)^{\dagger}\|_2 = \|\mathrm{diag}^{-1}(\sin\boldsymbol{\theta}(\mathbf{Y}_t, \mathbf{V}_q))\|_2 = \sin^{-1}\theta_{\min}(\mathbf{Y}_t, \mathbf{V}_q)$ and

$$\|\mathbf{E}_t\|_2 \le \|\begin{bmatrix}\boldsymbol{\xi}_t \\ \mathbf{0}\end{bmatrix}\|_2\|\mathbf{S}_t\|_2 \le \|\begin{bmatrix}\boldsymbol{\xi}_t \\ \mathbf{0}\end{bmatrix}\|_2 = \|\boldsymbol{\xi}_t\|_2,$$

one gets that

$$\begin{aligned}
\varphi_1 &\le \|\mathbf{\Gamma Q \Lambda}_{-q}(\mathbf{\Gamma Q})^{-1}\|_2 + \|\mathbf{\Gamma}\|_2\|\mathbf{\Gamma}^{-1}\|_2\|(\mathbf{V}_{-q}^{\mathrm{H}}\mathbf{Y}_t)^{\dagger}\|_2\|\boldsymbol{\xi}_t\|_2 \\
&\le \sqrt{\beta} + \frac{1+\beta}{1+\gamma} + \frac{1}{256}\frac{\sin\theta_{\max}(\mathbf{Y}_t, \mathbf{V}_q)}{\sin\theta_{\min}(\mathbf{Y}_t, \mathbf{V}_q)}\frac{(\lambda_k^+ - \sqrt{\beta})^2(1+\gamma)}{(1+l_1)(1+\beta)} \quad \text{(using Eq. (9) and assumption on } \|\boldsymbol{\xi}_t\|_2) \\
&\le \sqrt{\beta} + \frac{\lambda_k^+ - \sqrt{\beta}}{16} + \frac{\lambda_k^+ - \sqrt{\beta}}{16}. \qquad (\text{using}\frac{1+\beta}{1+\gamma} = \frac{\lambda_k^+ - \sqrt{\beta}}{16})
\end{aligned}$$

On the other hand,

$$\begin{aligned}
\varphi_2 &\le \|\mathbf{Z}_2(\mathbf{Z}_2 + \mathbf{\Lambda}_q^{-1}\mathbf{W}_q^{\mathrm{H}}\mathbf{E}_t)^{\top}((\mathbf{Z}_2 + \mathbf{\Lambda}_q^{-1}\mathbf{W}_q^{\mathrm{H}}\mathbf{E}_t)(\mathbf{Z}_2 + \mathbf{\Lambda}_q^{-1}\mathbf{W}_q^{\mathrm{H}}\mathbf{E}_t)^{\top})^{-1}\|_2 \\
&= \|(\mathbf{Z}_2\mathbf{Z}_2^{\top} + \mathbf{Z}_2(\mathbf{W}_q^{\mathrm{H}}\mathbf{E}_t)^{\top}\mathbf{\Lambda}_q^{-1})(\mathbf{Z}_2\mathbf{Z}_2^{\top} + \mathbf{Z}_2(\mathbf{W}_q^{\mathrm{H}}\mathbf{E}_t)^{\top}\mathbf{\Lambda}_q^{-1} + \\
&\qquad \mathbf{\Lambda}_q^{-1}\mathbf{W}_q^{\mathrm{H}}\mathbf{E}_t\mathbf{Z}_2^{\top} + \mathbf{\Lambda}_q^{-1}\mathbf{W}_q^{\mathrm{H}}\mathbf{E}_t(\mathbf{W}_q^{\mathrm{H}}\mathbf{E}_t)^{\top}\mathbf{\Lambda}_q^{-1})^{-1}\|_2 \\
&= \|(\mathbf{I} + \mathbf{Z}_2(\mathbf{W}_q^{\mathrm{H}}\mathbf{E}_t)^{\top}\mathbf{\Lambda}_q^{-1}(\mathbf{Z}_2\mathbf{Z}_2^{\top})^{-1})(\mathbf{I} + \mathbf{Z}_2(\mathbf{W}_q^{\mathrm{H}}\mathbf{E}_t)^{\top}\mathbf{\Lambda}_q^{-1}(\mathbf{Z}_2\mathbf{Z}_2^{\top})^{-1} + \\
&\qquad \mathbf{\Lambda}_q^{-1}\mathbf{W}_q^{\mathrm{H}}\mathbf{E}_t\mathbf{Z}_2^{\top}(\mathbf{Z}_2\mathbf{Z}_2^{\top})^{-1} + \mathbf{\Lambda}_q^{-1}\mathbf{W}_q^{\mathrm{H}}\mathbf{E}_t(\mathbf{W}_q^{\mathrm{H}}\mathbf{E}_t)^{\top}\mathbf{\Lambda}_q^{-1}(\mathbf{Z}_2\mathbf{Z}_2^{\top})^{-1})^{-1}\|_2 \\
&\le (1 + \|\mathbf{Z}_2(\mathbf{W}_q^{\mathrm{H}}\mathbf{E}_t)^{\top}\mathbf{\Lambda}_q^{-1}(\mathbf{Z}_2\mathbf{Z}_2^{\top})^{-1}\|_2)(1 - \|\mathbf{Z}_2(\mathbf{W}_q^{\mathrm{H}}\mathbf{E}_t)^{\top}\mathbf{\Lambda}_q^{-1}(\mathbf{Z}_2\mathbf{Z}_2^{\top})^{-1}\|_2 - \\
&\qquad \|\mathbf{\Lambda}_q^{-1}\mathbf{W}_q^{\mathrm{H}}\mathbf{E}_t\mathbf{Z}_2^{\top}(\mathbf{Z}_2\mathbf{Z}_2^{\top})^{-1}\|_2 - \|\mathbf{\Lambda}_q^{-1}\mathbf{W}_q^{\mathrm{H}}\mathbf{E}_t(\mathbf{W}_q^{\mathrm{H}}\mathbf{E}_t)^{\top}\mathbf{\Lambda}_q^{-1}(\mathbf{Z}_2\mathbf{Z}_2^{\top})^{-1})^{-1}\|_2)^{-1}. \\
&\le (1 + \|\mathbf{Z}_2\|_2\|\mathbf{W}_q^{\mathrm{H}}\mathbf{E}_t\|_2\|\mathbf{\Lambda}_q^{-1}\|_2\|(\mathbf{Z}_2\mathbf{Z}_2^{\top})^{-1}\|_2)(1 - \|\mathbf{Z}_2\|_2\|\mathbf{W}_q^{\mathrm{H}}\mathbf{E}_t\|_2\|\mathbf{\Lambda}_q^{-1}\|_2\|(\mathbf{Z}_2\mathbf{Z}_2^{\top})^{-1}\|_2 \\
&\qquad -\|\mathbf{\Lambda}_q^{-1}\|_2\|\mathbf{W}_q^{\mathrm{H}}\mathbf{E}_t\|_2\|\mathbf{Z}_2^{\top}(\mathbf{Z}_2\mathbf{Z}_2^{\top})^{-1}\|_2 - \|\mathbf{\Lambda}_q^{-1}\|_2^2\|\mathbf{W}_q^{\mathrm{H}}\mathbf{E}_t\|_2^2\|(\mathbf{Z}_2\mathbf{Z}_2^{\top})^{-1})^{-1}\|_2)^{-1}.
\end{aligned}$$

Observe that $\|\mathbf{Z}_2\|_2 = \cos\theta_{\min}(\mathbf{Y}_t, \mathbf{W}_q)$, $\|(\mathbf{Z}_2\mathbf{Z}_2^{\top})^{-1}\|_2 = \cos^{-2}\theta_{\max}(\mathbf{Y}_t, \mathbf{W}_q)$, and

$$\|\mathbf{Z}_2^{\top}(\mathbf{Z}_2\mathbf{Z}_2^{\top})^{-1}\|_2 = \cos^{-1}\theta_{\max}(\mathbf{Y}_t, \mathbf{W}_q).$$

Particularly, it holds by Eq. (7) that

$$\begin{aligned}\|\mathbf{W}_q^{\mathrm{H}}\mathbf{E}_t\|_2 &= \|\mathbf{E}_t^{\mathrm{H}}\mathbf{W}_q\|_2 = \|\mathbf{S}_t\begin{bmatrix}\boldsymbol{\xi}_t\\ \mathbf{0}\end{bmatrix}^\top\begin{bmatrix}\mathbf{U}_q\mathbf{D}_q(\beta)\\ -\mathbf{U}_q\mathbf{K}_q(\beta)\end{bmatrix}\|_2 = \|\mathbf{S}_t\boldsymbol{\xi}_t^\top\mathbf{U}_q\mathbf{D}_q(\beta)\|_2\\ &\le \|\mathbf{S}_t\|_2\|\boldsymbol{\xi}_t^\top\mathbf{U}_q\|_2\|\mathbf{D}_q(\beta)\|_2 \le \|\mathbf{U}_q^\top\boldsymbol{\xi}_t\|_2.\end{aligned}$$

We then get that

$$\begin{aligned}\varphi_2 \le\ & (1+\frac{\cos\theta_{\min}(\mathbf{Y}_t,\mathbf{W}_q)}{\cos^2\theta_{\max}(\mathbf{Y}_t,\mathbf{W}_q)}\frac{\|\mathbf{U}_q^\top\boldsymbol{\xi}_t\|_2}{\lambda_q^+})(1-\frac{\cos\theta_{\min}(\mathbf{Y}_t,\mathbf{W}_q)}{\cos^2\theta_{\max}(\mathbf{Y}_t,\mathbf{W}_q)}\frac{\|\mathbf{U}_q^\top\boldsymbol{\xi}_t\|_2}{\lambda_q^+}\\ &-\frac{\|\mathbf{U}_q^\top\boldsymbol{\xi}_t\|_2}{\lambda_q^+\cos\theta_{\max}(\mathbf{Y}_t,\mathbf{W}_q)}-(\frac{\|\mathbf{U}_q^\top\boldsymbol{\xi}_t\|_2}{\lambda_q^+\cos\theta_{\max}(\mathbf{Y}_t,\mathbf{W}_q)})^2)^{-1},\end{aligned}$$

Further by the assumption on $\|\mathbf{U}_q^\top\boldsymbol{\xi}_t\|_2$, it holds that

$$\begin{aligned}\varphi_2 &\le (1+\frac{l_2(\lambda_k^+-\sqrt{\beta})}{16(1+l_2)(1+\lambda_k^+)})(1-\frac{l_2(\lambda_k^+-\sqrt{\beta})}{16(1+l_2)(1+\lambda_k^+)}\\ &\quad -\frac{\lambda_k^+-\sqrt{\beta}}{16(1+l_2)(1+\lambda_k^+)}-(\frac{\lambda_k^+-\sqrt{\beta}}{16(1+l_2)(1+\lambda_k^+)})^2)^{-1}\\ &\le (1+\frac{\lambda_k^+-\sqrt{\beta}}{16(1+\lambda_k^+)})(1-\frac{3(\lambda_k^+-\sqrt{\beta})}{16(1+\lambda_k^+)})^{-1}.\end{aligned}$$

We now can write that

$$\begin{aligned}\|\mathbf{H}_{t+1}\|_2 &\le \varphi_1\varphi_2\|\mathbf{H}_t\|_2\|\mathbf{\Lambda}_k^{-1}\|_2\\ &\le (\sqrt{\beta}+\frac{\lambda_k^+-\sqrt{\beta}}{8})(1+\frac{\lambda_k^+-\sqrt{\beta}}{16(1+\lambda_k^+)})(\lambda_k^+-\frac{3\lambda_k^+(\lambda_k^+-\sqrt{\beta})}{16(1+\lambda_k^+)})^{-1}\|\mathbf{H}_t\|_2\\ &= \frac{\sqrt{\beta}+\frac{\lambda_k^+-\sqrt{\beta}}{8}+(\sqrt{\beta}+\frac{\lambda_k^+-\sqrt{\beta}}{8})\frac{\lambda_k^+-\sqrt{\beta}}{16(1+\lambda_k^+)}}{\lambda_k^+-\frac{3\lambda_k^+(\lambda_k^+-\sqrt{\beta})}{16(1+\lambda_k^+)}}\|\mathbf{H}_t\|_2\\ &\le \frac{\sqrt{\beta}+\frac{\lambda_k^+-\sqrt{\beta}}{8}+\frac{\lambda_k^+-\sqrt{\beta}}{8}}{\lambda_k^+-\frac{3\lambda_k^+(\lambda_k^+-\sqrt{\beta})}{16(1+\lambda_k^+)}}\|\mathbf{H}_t\|_2 \qquad (\text{using } \sqrt{\beta}<\lambda_q^+\le\lambda_k^+)\\ &\le \frac{\sqrt{\beta}+\frac{\lambda_k^+-\sqrt{\beta}}{4}}{\lambda_k^+-\frac{\lambda_k^+-\sqrt{\beta}}{4}}\|\mathbf{H}_t\|_2 = \frac{\lambda_k^++3\sqrt{\beta}}{3\lambda_k^++\sqrt{\beta}}\|\mathbf{H}_t\|_2 = (1-\frac{2(\lambda_k^+-\sqrt{\beta})}{3\lambda_k^++\sqrt{\beta}})\|\mathbf{H}_t\|_2\\ &\le (1-\frac{\lambda_k^+-\sqrt{\beta}}{2\lambda_k^+})\|\mathbf{H}_t\|_2,\end{aligned}$$

where

$$\frac{\lambda_k^+ - \sqrt{\beta}}{\lambda_k^+} = \frac{\lambda_k - 2\sqrt{\beta} + \sqrt{\lambda_k^2 - 4\beta}}{\lambda_k + \sqrt{\lambda_k^2 - 4\beta}} = \frac{\sqrt{\lambda_k - 2\sqrt{\beta}}(\sqrt{\lambda_k - 2\sqrt{\beta}} + \sqrt{\lambda_k + 2\sqrt{\beta}}}{\lambda_k + \sqrt{\lambda_k^2 - 4\beta}}$$
$$> \frac{\sqrt{\lambda_k - 2\sqrt{\beta}}\sqrt{\lambda_k}}{2\lambda_k} = \frac{1}{2}\sqrt{\frac{\lambda_k - 2\sqrt{\beta}}{\lambda_k}}.$$

Therefore, it holds that

$$\Phi_{t+1} \leq (1 - \frac{1}{4}\sqrt{\frac{\lambda_k - 2\sqrt{\beta}}{\lambda_k}}\,)\Phi_t.$$

□

Lemma 5 $\sin\theta_{\max}(\mathbf{Y}_t, \mathbf{V}_k) = \|(\mathbf{I} - \mathbf{Y}_t\mathbf{Y}_t^\top)\mathbf{V}_k\|_2 \leq (1 + 2\frac{1+\beta}{\sqrt{\lambda_q^2 - 4\beta}})\Phi_t.$

Proof Let $\mathbf{\Xi}_k = \mathrm{diag}(\frac{1}{\sqrt{1+\alpha_1^2}}, \cdots, \frac{1}{\sqrt{1+\alpha_k^2}})$ with $\alpha_j = \frac{1+\beta}{\sqrt{\lambda_j^2 - 4\beta}}$, i.e., $\mathbf{\Xi}_k = \begin{bmatrix}\mathbf{I}_k & \mathbf{0}\end{bmatrix}(\mathbf{I} + \mathbf{\Omega}\mathbf{\Omega}^{\mathrm{H}})^{-\frac{1}{2}}\begin{bmatrix}\mathbf{I}_k \\ \mathbf{0}\end{bmatrix}$ is the $k \times k$ leading principal submatrix of $(\mathbf{I} + \mathbf{\Omega}\mathbf{\Omega}^{\mathrm{H}})^{-\frac{1}{2}}$. We then have that

$$\begin{aligned}
\|(\mathbf{I} - \mathbf{Y}_t\mathbf{Y}_t^\top)\mathbf{V}_k\|_2 &= \|(\mathbf{I} - \mathbf{Y}_t\mathbf{Y}_t^\top)\mathbf{V}_k\mathbf{\Xi}_k^{-1}\mathbf{\Xi}_k\|_2 \leq \|(\mathbf{I} - \mathbf{Y}_t\mathbf{Y}_t^\top)\mathbf{V}_k\mathbf{\Xi}_k^{-1}\|_2\|\mathbf{\Xi}_k\|_2 \\
&\leq \|(\mathbf{I} - \mathbf{Y}_t\mathbf{Y}_t^\top)\mathbf{V}_k\mathbf{\Xi}_k^{-1}\|_2 = \max_{\mathbf{x}\in\mathbb{R}^{k\times 1}:\|\mathbf{x}\|_2=1}\|(\mathbf{I} - \mathbf{Y}_t\mathbf{Y}_t^\top)\mathbf{V}_k\mathbf{\Xi}_k^{-1}\mathbf{x}\|_2 \\
&= \max_{\mathbf{x}\in\mathbb{R}^{k\times 1}:\|\mathbf{x}\|_2=1}\|\mathbf{V}_k\mathbf{\Xi}_k^{-1}\mathbf{x} - \mathbf{Y}_t\mathbf{Y}_t^\top\mathbf{V}_k\mathbf{\Xi}_k^{-1}\mathbf{x}\|_2 \\
&\leq \max_{\mathbf{x}\in\mathbb{R}^{k\times 1}:\|\mathbf{x}\|_2=1}\|\mathbf{V}_k\mathbf{\Xi}_k^{-1}\mathbf{x} - \mathbf{Y}_t(\mathbf{W}_q^{\mathrm{H}}\mathbf{Y}_t)^\dagger\begin{bmatrix}\mathbf{I}_k \\ \mathbf{0}\end{bmatrix}\mathbf{x}\|_2 \\
&= \max_{\mathbf{x}\in\mathbb{R}^{k\times 1}:\|\mathbf{x}\|_2=1}\|(\mathbf{V}_k\mathbf{\Xi}_k^{-1} - \mathbf{Y}_t(\mathbf{W}_q^{\mathrm{H}}\mathbf{Y}_t)^\dagger\begin{bmatrix}\mathbf{I}_k \\ \mathbf{0}\end{bmatrix})\mathbf{x}\|_2 \\
&= \|\mathbf{V}_k\mathbf{\Xi}_k^{-1} - \mathbf{Y}_t(\mathbf{W}_q^{\mathrm{H}}\mathbf{Y}_t)^\dagger\begin{bmatrix}\mathbf{I}_k \\ \mathbf{0}\end{bmatrix}\|_2,
\end{aligned}$$

where the second and last equalities are by the definition of matrix 2-norm, and the last inequality is by the definition of vector projection onto a column space of a matrix (see the second proof of Lemma 2.3 or A.4 in Balcan et al. (2016)). To proceed, we can write that

$$\begin{aligned}
\|(\mathbf{I} - \mathbf{Y}_t\mathbf{Y}_t^\top)\mathbf{V}_k\|_2 &\leq \|\mathbf{V}_k\mathbf{\Xi}_k^{-1} - \mathbf{Y}_t(\mathbf{W}_q^{\mathrm{H}}\mathbf{Y}_t)^\dagger\begin{bmatrix}\mathbf{I}_k \\ \mathbf{0}\end{bmatrix}\|_2 \\
&= \|\begin{bmatrix}\mathbf{W}_q & \mathbf{V}_{-q}\end{bmatrix}^{-\mathrm{H}}\begin{bmatrix}\mathbf{W}_q & \mathbf{V}_{-q}\end{bmatrix}^{\mathrm{H}}(\mathbf{V}_k\mathbf{\Xi}_k^{-1} - \mathbf{Y}_t(\mathbf{W}_q^{\mathrm{H}}\mathbf{Y}_t)^\dagger\begin{bmatrix}\mathbf{I}_k \\ \mathbf{0}\end{bmatrix})\|_2 \\
&\leq \|\begin{bmatrix}\mathbf{W}_q & \mathbf{V}_{-q}\end{bmatrix}^{-\mathrm{H}}\|_2\|\begin{bmatrix}\mathbf{W}_q & \mathbf{V}_{-q}\end{bmatrix}^{\mathrm{H}}(\mathbf{V}_k\mathbf{\Xi}_k^{-1} - \mathbf{Y}_t(\mathbf{W}_q^{\mathrm{H}}\mathbf{Y}_t)^\dagger\begin{bmatrix}\mathbf{I}_k \\ \mathbf{0}\end{bmatrix})\|_2 \\
&= \frac{1}{\sigma_{\min}(\begin{bmatrix}\mathbf{W}_q & \mathbf{V}_{-q}\end{bmatrix})}\|\begin{bmatrix}\mathbf{W}_q & \mathbf{V}_{-q}\end{bmatrix}^{\mathrm{H}}(\mathbf{V}_k\mathbf{\Xi}_k^{-1} - \mathbf{Y}_t(\mathbf{W}_q^{\mathrm{H}}\mathbf{Y}_t)^\dagger\begin{bmatrix}\mathbf{I}_k \\ \mathbf{0}\end{bmatrix})\|_2.
\end{aligned}$$

Noting that $\mathbf{W}_q = (\mathbf{V}_q - \mathbf{V}_{-q}\mathbf{\Omega}^{\mathrm{H}})(\mathrm{I} + \mathbf{\Omega}\mathbf{\Omega}^{\mathrm{H}})^{-\frac{1}{2}}$, it holds that

$$\mathbf{W}_q^{\mathrm{H}}\mathbf{V}_k\mathbf{\Xi}_k^{-1} = (\mathrm{I} + \mathbf{\Omega}\mathbf{\Omega}^{\mathrm{H}})^{-\frac{1}{2}}\mathbf{V}_q^{\mathrm{H}}\mathbf{V}_k\mathbf{\Xi}_k^{-1} = (\mathrm{I} + \mathbf{\Omega}\mathbf{\Omega}^{\mathrm{H}})^{-\frac{1}{2}}\begin{bmatrix}\mathbf{I}_k\\ \mathbf{0}\end{bmatrix}\mathbf{\Xi}_k^{-1} = \begin{bmatrix}\mathbf{\Xi}_k\\ \mathbf{0}\end{bmatrix}\mathbf{\Xi}_k^{-1} = \begin{bmatrix}\mathbf{I}_k\\ \mathbf{0}\end{bmatrix}.$$

Thus, we have that

$$\begin{aligned}
& \|\begin{bmatrix}\mathbf{W}_q & \mathbf{V}_{-q}\end{bmatrix}^{\mathrm{H}}(\mathbf{V}_k\mathbf{\Xi}_k^{-1} - \mathbf{Y}_t(\mathbf{W}_q^{\mathrm{H}}\mathbf{Y}_t)^{\dagger}\begin{bmatrix}\mathbf{I}_k\\ \mathbf{0}\end{bmatrix})\|_2 \\
\leq\ & \|\mathbf{W}_q^{\mathrm{H}}\mathbf{V}_k\mathbf{\Xi}_k^{-1} - (\mathbf{W}_q^{\mathrm{H}}\mathbf{Y}_t)(\mathbf{W}_q^{\mathrm{H}}\mathbf{Y}_t)^{\dagger}\begin{bmatrix}\mathbf{I}_k\\ \mathbf{0}\end{bmatrix}\|_2 + \|\mathbf{V}_{-q}^{\mathrm{H}}\mathbf{V}_k\mathbf{\Xi}_k^{-1} - \mathbf{V}_{-q}^{\mathrm{H}}\mathbf{Y}_t(\mathbf{W}_q^{\mathrm{H}}\mathbf{Y}_t)^{\dagger}\begin{bmatrix}\mathbf{I}_k\\ \mathbf{0}\end{bmatrix}\|_2 \\
=\ & \|\mathbf{V}_{-q}^{\mathrm{H}}\mathbf{Y}_t(\mathbf{W}_q^{\mathrm{H}}\mathbf{Y}_t)^{\dagger}\begin{bmatrix}\mathbf{I}_k\\ \mathbf{0}\end{bmatrix}\|_2 = \|(\mathbf{\Gamma}\mathbf{Q})^{-1}(\mathbf{\Gamma}\mathbf{Q}\mathbf{V}_{-q}^{\mathrm{H}}\mathbf{Y}_t)(\mathbf{W}_q^{\mathrm{H}}\mathbf{Y}_t)^{\dagger}\begin{bmatrix}\mathbf{I}_k\\ \mathbf{0}\end{bmatrix}\|_2 \\
\leq\ & \|(\mathbf{\Gamma}\mathbf{Q})^{-1}\|_2\|(\mathbf{\Gamma}\mathbf{Q}\mathbf{V}_{-q}^{\mathrm{H}}\mathbf{Y}_t)(\mathbf{W}_q^{\mathrm{H}}\mathbf{Y}_t)^{\dagger}\begin{bmatrix}\mathbf{I}_k\\ \mathbf{0}\end{bmatrix}\|_2 \leq \|(\mathbf{\Gamma}\mathbf{Q}\mathbf{V}_{-q}^{\mathrm{H}}\mathbf{Y}_t)(\mathbf{W}_q^{\mathrm{H}}\mathbf{Y}_t)^{\dagger}\begin{bmatrix}\mathbf{I}_k\\ \mathbf{0}\end{bmatrix}\|_2 = \Phi_t,
\end{aligned}$$

arriving at

$$\|(\mathbf{I} - \mathbf{Y}_t\mathbf{Y}_t^{\top})\mathbf{V}_k\|_2 \leq \frac{\Phi_t}{\sigma_{\min}(\begin{bmatrix}\mathbf{W}_q & \mathbf{V}_{-q}\end{bmatrix})}.$$

It remains to figure out $\sigma_{\min}(\begin{bmatrix}\mathbf{W}_q & \mathbf{V}_{-q}\end{bmatrix})$ as follows.

$$\begin{aligned}
\sigma_{\min}^2(\begin{bmatrix}\mathbf{W}_q & \mathbf{V}_{-q}\end{bmatrix}) &= \lambda_{\min}(\begin{bmatrix}\mathbf{W}_q & \mathbf{V}_{-q}\end{bmatrix}^{\mathrm{H}}\begin{bmatrix}\mathbf{W}_q & \mathbf{V}_{-q}\end{bmatrix}) = \lambda_{\min}(\begin{bmatrix}\mathbf{I} & \mathbf{W}_q^{\mathrm{H}}\mathbf{V}_{-q}\\ \mathbf{V}_{-q}^{\mathrm{H}}\mathbf{W}_q & \mathbf{I}\end{bmatrix}) \\
&= \lambda_{\min}(\begin{bmatrix}\mathbf{I} & -(\mathbf{I} + \mathbf{\Omega}\mathbf{\Omega}^{\mathrm{H}})^{-\frac{1}{2}}\mathbf{\Omega}\\ -((\mathbf{I} + \mathbf{\Omega}\mathbf{\Omega}^{\mathrm{H}})^{-\frac{1}{2}}\mathbf{\Omega})^{\top} & \mathbf{I}\end{bmatrix}) \\
&= \lambda_{\min}(\mathrm{diag}(\begin{bmatrix}1 & -\frac{\alpha_1}{\sqrt{1+\alpha_1^2}}\\ -\frac{\alpha_1}{\sqrt{1+\alpha_1^2}} & 1\end{bmatrix}, \cdots, \begin{bmatrix}1 & -\frac{\alpha_q}{\sqrt{1+\alpha_q^2}}\\ -\frac{\alpha_q}{\sqrt{1+\alpha_q^2}} & 1\end{bmatrix}, \mathbf{I}_{2(n-q)})) \\
&= \min_{1\leq j\leq q}\lambda_{\min}(\begin{bmatrix}1 & -\frac{\alpha_j}{\sqrt{1+\alpha_j^2}}\\ -\frac{\alpha_j}{\sqrt{1+\alpha_j^2}} & 1\end{bmatrix}),
\end{aligned}$$

where $\alpha_j = \frac{1+\beta}{\sqrt{\lambda_j^2 - 4\beta}}$, and the second last equality has used Lemma 3 and orthogonal invariance of eigenvalues (the left and right permutation matrices we have used here are orthogonal). The eigenvalues of $\begin{bmatrix}1 & -\frac{\alpha_j}{\sqrt{1+\alpha_j^2}}\\ -\frac{\alpha_j}{\sqrt{1+\alpha_j^2}} & 1\end{bmatrix}$ are $1 \pm \frac{\alpha_j}{\sqrt{1+\alpha_j^2}}$ which are the roots of the following equation:

$$x^2 - \mathrm{tr}(\begin{bmatrix}1 & -\frac{\alpha_j}{\sqrt{1+\alpha_j^2}}\\ -\frac{\alpha_j}{\sqrt{1+\alpha_j^2}} & 1\end{bmatrix})x + \det(\begin{bmatrix}1 & -\frac{\alpha_j}{\sqrt{1+\alpha_j^2}}\\ -\frac{\alpha_j}{\sqrt{1+\alpha_j^2}} & 1\end{bmatrix}) = 0, \text{ i.e., } x^2 - 2x + \frac{1}{1+\alpha_j^2} = 0.$$

Thus, we have that

$$\begin{aligned}
\sigma_{\min}^2([\mathbf{W}_q \quad \mathbf{V}_{-q}]) &= 1-\frac{\alpha_q}{\sqrt{1+\alpha_q^2}} = 1-\frac{1+\beta}{\sqrt{\lambda_q^2-4\beta+(1+\beta)^2}} \\
&= \frac{\sqrt{\lambda_q^2-4\beta+(1+\beta)^2}-(1+\beta)}{\sqrt{\lambda_q^2-4\beta+(1+\beta)^2}} \\
&= \frac{\lambda_q^2-4\beta+(1+\beta)^2-(1+\beta)^2}{\sqrt{\lambda_q^2-4\beta+(1+\beta)^2}(\sqrt{\lambda_q^2-4\beta+(1+\beta)^2}+1+\beta)} \\
&= \frac{\lambda_q^2-4\beta}{\sqrt{\lambda_q^2-4\beta+(1+\beta)^2}(\sqrt{\lambda_q^2-4\beta+(1+\beta)^2}+1+\beta)} \\
&\geq \frac{\lambda_q^2-4\beta}{(\sqrt{\lambda_q^2-4\beta+(1+\beta)^2}+1+\beta)^2} \geq \frac{\lambda_q^2-4\beta}{(\sqrt{\lambda_q^2-4\beta}+2(1+\beta))^2}.
\end{aligned}$$

Therefore, it holds that

$$\|(\mathbf{I}-\mathbf{Y}_t\mathbf{Y}_t^\top)\mathbf{V}_k\|_2 \leq (1+2\frac{1+\beta}{\sqrt{\lambda_q^2-4\beta}})\Phi_t.$$

□

Lemma 6 If $\sin\theta_{\max}(\mathbf{Y}_t,\mathbf{V}_k) < \frac{\lambda_k^+}{k\sqrt{1+(\lambda_k^+)^2}}\epsilon$, then $\sin\theta_{\max}(\mathbf{X}_t,\mathbf{U}_k) < \epsilon$.

Proof Since we now have that

$$\begin{aligned}
k-\|\mathbf{V}_k^\top\mathbf{Y}_t\|_F^2 &= k-\|\mathrm{diag}(\cos\theta_1(\mathbf{Y}_t,\mathbf{V}_k),\cdots,\cos\theta_k(\mathbf{Y}_t,\mathbf{V}_k))\|_F^2 \\
&= \sum_{j=1}^k \sin^2\theta_j(\mathbf{Y}_t,\mathbf{V}_k) \leq k\sin^2\theta_{\max}(\mathbf{Y}_t,\mathbf{V}_k) < \frac{(\lambda_k^+)^2\epsilon^2}{k(1+(\lambda_k^+)^2)},
\end{aligned}$$

it holds that $\|\mathbf{V}_k^\top\mathbf{Y}_t\|_F^2 > k-\frac{(\lambda_k^+)^2\epsilon^2}{k(1+(\lambda_k^+)^2)}$. Recall that $(\lambda_j,\mathbf{u}_j)$ is an eigenpair of $\mathbf{A}$ indexed in descending order of its eigenvalues, $\mathbf{U}_k=[\mathbf{u}_1 \quad \cdots \quad \mathbf{u}_k]$, and $\mathbf{V}_k=\begin{bmatrix}\mathbf{U}_k\mathbf{D}_k(1)\\ \mathbf{U}_k\mathbf{K}_k(1)\end{bmatrix}$ (see notations in Lemma 2, Eq. (4) as well as the paragraph right before Lemma 5). Let $\mathbf{v}_j$ be the j-th column of $\mathbf{V}_k$, i.e., $\begin{bmatrix}a_j\mathbf{u}_j\\ b_j\mathbf{u}_j\end{bmatrix}$, where $a_j=\frac{\lambda_j^+}{\sqrt{1+|\lambda_j^+|^2}}$ and $b_j=\frac{1}{\sqrt{1+|\lambda_j^+|^2}}$, for $j=1,\cdots,k$. We then have for any

$j = 1, \cdots, k$ that

$$\begin{aligned}
\|\mathbf{v}_j^\top \mathbf{Y}_t\|_2^2 &= \| \begin{bmatrix} a_j \mathbf{u}_j \\ b_j \mathbf{u}_j \end{bmatrix}^\top \begin{bmatrix} \mathbf{X}_t \\ \mathbf{X}_{t-1}\mathbf{R}_t^{-1} \end{bmatrix} (\mathbf{I} + \mathbf{R}_t^{-\top}\mathbf{R}_t^{-1})^{-\frac{1}{2}} \|_2^2 \\
&= \| \begin{bmatrix} a_j \mathbf{X}_t^\top \mathbf{u}_j \\ b_j \mathbf{X}_{t-1}^\top \mathbf{u}_j \end{bmatrix}^\top \begin{bmatrix} \mathbf{I} \\ \mathbf{R}_t^{-1} \end{bmatrix} (\mathbf{I} + \mathbf{R}_t^{-\top}\mathbf{R}_t^{-1})^{-\frac{1}{2}} \|_2^2 \\
&\leq \| \begin{bmatrix} a_j \mathbf{X}_t^\top \mathbf{u}_j \\ b_j \mathbf{X}_{t-1}^\top \mathbf{u}_j \end{bmatrix}^\top \|_2^2 \| \begin{bmatrix} \mathbf{I} \\ \mathbf{R}_t^{-1} \end{bmatrix} (\mathbf{I} + \mathbf{R}_t^{-\top}\mathbf{R}_t^{-1})^{-\frac{1}{2}} \|_2^2 = \| \begin{bmatrix} a_j \mathbf{X}_t^\top \mathbf{u}_j \\ b_j \mathbf{X}_{t-1}^\top \mathbf{u}_j \end{bmatrix}^\top \|_2^2 \\
&= a_j^2 \|\mathbf{X}_t^\top \mathbf{u}_j\|_2^2 + b_j^2 \|\mathbf{X}_{t-1}^\top \mathbf{u}_j\|_2^2,
\end{aligned}$$

where the second last equality is due to that $\begin{bmatrix} \mathbf{I} \\ \mathbf{R}_t^{-1} \end{bmatrix} (\mathbf{I} + \mathbf{R}_t^{-\top}\mathbf{R}_t^{-1})^{-\frac{1}{2}}$ is orthonormal. Thus, if there exists a certain $j \in \{1, \cdots, k\}$ such that $\|\mathbf{X}_t^\top \mathbf{u}_j\|_2^2 \leq 1 - \frac{\epsilon^2}{k}$, noting that $\|\mathbf{X}_{t-1}^\top \mathbf{u}_j\|_2 \leq 1$ we then would have that

$$\|\mathbf{v}_j^\top \mathbf{Y}_t\|_2^2 \leq a_j^2(1 - \frac{\epsilon^2}{k}) + b_j^2 = (a_j^2 + b_j^2) - a_j^2 \frac{\epsilon^2}{k} \leq 1 - a_k^2 \frac{\epsilon^2}{k}.$$

Consequently, it would hold that

$$\|\mathbf{V}_k^\top \mathbf{Y}_t\|_F^2 = \sum_{s=1}^k \|\mathbf{v}_s^\top \mathbf{Y}_t\|_2^2 = \|\mathbf{v}_j^\top \mathbf{Y}_t\|_2^2 + \sum_{\substack{s=1 \\ s \neq j}}^k \|\mathbf{v}_s^\top \mathbf{Y}_t\|_2^2 \leq 1 - a_k^2 \frac{\epsilon^2}{k} + k - 1 = k - a_k^2 \frac{\epsilon^2}{k},$$

which contradicts the fact at the beginning that $\|\mathbf{V}_k^\top \mathbf{Y}_t\|_F^2 > k - a_k^2 \frac{\epsilon^2}{k}$. Therefore, it must hold for any $j = 1, \cdots, k$ that $\|\mathbf{u}_j^\top \mathbf{X}_t\|_2^2 > 1 - \frac{\epsilon^2}{k}$, and thus $\|\mathbf{U}_k^\top \mathbf{X}_t\|_F^2 = \sum_{j=1}^k \|\mathbf{u}_j^\top \mathbf{X}_t\|_2^2 > k - \epsilon^2$. Finally, we get that

$$\begin{aligned}
\sin\theta_{\max}(\mathbf{X}_t, \mathbf{U}_k) &\leq \|\mathrm{diag}(\sin\theta_1(\mathbf{X}_t, \mathbf{U}_k), \cdots, \sin\theta_k(\mathbf{X}_t, \mathbf{U}_k))\|_F \\
&= \sqrt{k - \|\mathbf{U}_k^\top \mathbf{X}_t\|_F^2} < \epsilon.
\end{aligned}$$

□