4-2023

On the Accelerated Noise-Tolerant Power Method

Zhiqiang Xu
Mohamed Bin Zayed University of Artificial Intelligence

Recommended Citation

Z. Xu, "On the Accelerated Noise-Tolerant Power Method", in "26th Intl. Conf. on Artificial Intelligence and Statistics (AISTATS 2023), vol. 206, pp. 7147-7175, Apr 2023. Available at: https://proceedings.mlr.press/v206/xu23g.html

On the Accelerated Noise-Tolerant Power Method

Zhiqiang Xu
Machine Learning Department
Mohamed bin Zayed University of Artificial Intelligence
Abu Dhabi, UAE
zhiqiang.xu@mbzuai.ac.ae

Abstract

We revisit the acceleration of the noise-tolerant power method for which, despite previous studies, the results remain unsatisfactory as they are either wrong or suboptimal, also lacking generality. In this work, we present a simple yet general and optimal analysis via noise-corrupted Chebyshev polynomials, which allows a larger iteration rank p than the target rank k, requires less noise conditions in a new form, and achieves the optimal iteration complexity $\Theta\left(\sqrt{\frac{\lambda_k-\lambda_{q+1}}{\lambda_k}}\log\frac{1}{\epsilon}\right)$ for some q satisfying $k \leq q \leq p$ in a certain regime of the momentum parameter. Interestingly, it shows dynamic dependence of the noise tolerance on the spectral gap, i.e., from linear at the beginning to square-root near convergence, while remaining commensurate with the previous in terms of overall tolerance. We relate our new form of noise norm conditions to the existing trigonometric one, which enables an improved analysis of generalized eigenspace computation and canonical correlation analysis. We conduct an extensive experimental study to showcase the great performance of the considered algorithm with a larger iteration rank $p > k$ across different applications.

1 INTRODUCTION

Power method is a classic algorithm for the dominant eigenspace computation required in many problems of machine learning and statistics. Recently, it has been an emerging trend to demand noisy per-iteration updates caused by such factors as privacy constraint, missing entries, sampling error, adversarial attack, and approximation error (Hardt and Roth, 2013; Mitliagkas et al., 2013; Liang et al., 2014; Liu et al., 2015; Ge et al., 2016). For example, for the privacy concern, we may want to add noises to each power iteration to avoid the data privacy leakage. Also, for the generalized eigenspace computation which involves the inverse of a large matrix, we may approximate the multiplication of the inverse matrix with vectors to avoid the costly matrix inversion for efficiency. This setting is common in downstream machine learning applications such as principal component analysis and canonical correlation analysis. The noisy power method (Hardt and Price, 2014) is a meta algorithm for this purpose that can handle varieties of noises $\boldsymbol{\xi}$ during power iterations. Both its convergence rate and noise conditions (indicating noise tolerance level) linearly depend on the consecutive spectral gap $(\lambda_k - \lambda_{k+1})$, where λ_i represents the i-th largest eigenvalue of the input real symmetric matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ and k is the target rank when a top-k eigenspace of $\mathbf{A}$, denoted as $\mathbf{U}_k$, is desired. Acceleration, as an appealing feature, has been considered for the noisy power method as well, via the use of a larger iteration rank p than target rank k (Balcan et al., 2016), or the momentum under $p = k$ (Mai and Johansson, 2019), or the momentum under $p \geq k$ (Xu and Li, 2022). A larger iteration rank leads to dependence on a non-consecutive spectral gap which is an enlarged spectral gap compared to the consecutive one, while momentum yields square-root dependence on spectral gap. However, these results are still unsatisfactory. Balcan et al. (2016); Mai and Johansson (2019) considered only a single type of acceleration. Unfortunately, Mai and Johansson (2019) gave a wrong analysis[1] which resulted in a wrong noise tolerance bound that scales with $\sqrt{\lambda_k - \lambda_{k+1}}$ across iterations. This contradicts the folklore that accelerated methods have no better noise tolerance than their unaccelerated counterparts (Hardt and Roth, 2013). Also, it lacks generality as only a special kind of noise was discussed. Xu and Li (2022) combined both types of acceleration, but only achieves a sub-optimal iteration complexity

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206.

[1] See Section A of the Appendix for details.

which has an extra logarithmic factor on the spectral gap, and it lacks generality as well because it requires not two as usual but three noise conditions, one of which is quite restrictive for each iteration and thus limiting its applicability.

In this work, we present a simple yet general and optimal analysis for the accelerated noise-tolerant power method with momentum under $p \geq k$ for $t \geq 1$:

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

where $\beta > 0$ is the momentum parameter, $\boldsymbol{\xi}_t$ is the noise matrix for iteration $t \geq 0$, and $\mathbf{X}_{t+1}\mathbf{R}_{t+1}$ represents the QR factorization (Golub and Van Loan, 2013) of the right-hand side that can keep $\mathbf{X}_{t+1} \in \mathbb{R}^{n\times p}$ column-orthonormal and $\mathbf{R}_{t+1} \in \mathbb{R}^{p\times p}$ for all $t \geq 1$. Since we use the scaled Chebyshev acceleration (Xu et al., 2018), we start the iteration by setting $\mathbf{X}_0, \mathbf{X}_1 \in \mathbb{R}^{n\times p}$ to be Q-factor matrices of the QR factorization of an entry-wise i.i.d. standard Gaussian matrix $\mathbf{G} \in \mathbb{R}^{n\times p}$ and $\frac{1}{2}\mathbf{A}\mathbf{X}_0 + \boldsymbol{\xi}_0$, respectively, i.e.,

$$\mathbf{X}_0\mathbf{R}_0 = \mathbf{G}, \quad \mathbf{X}_1\mathbf{R}_1 = \frac{1}{2}\mathbf{A}\mathbf{X}_0 + \boldsymbol{\xi}_0 \in \mathbb{R}^{n\times p}. \quad (2)$$

To analyze the noise tolerance and convergence of Eq. (1), our first key step is to establish the connection to its non-orthonormal version:

$$\widehat{\mathbf{X}}_{t+1} = \mathbf{A}\widehat{\mathbf{X}}_t - \beta\widehat{\mathbf{X}}_{t-1} + \widehat{\boldsymbol{\xi}}_t \in \mathbb{R}^{n\times p}, \quad (3)$$

with initials

$$\widehat{\mathbf{X}}_0 = \mathbf{X}_0, \quad \widehat{\mathbf{X}}_1 = \frac{1}{2}\mathbf{A}\widehat{\mathbf{X}}_0 + \widehat{\boldsymbol{\xi}}_0, \quad (4)$$

where both $\widehat{\mathbf{X}}_t$ and $\widehat{\boldsymbol{\xi}}_t$ are unnormalized and thus may explode or vanish in a certain norm as the iteration proceeds. Thus, Eq. (3)-(4) are only used for theoretical analysis rather than practice. The second key step is to have $\widehat{\mathbf{X}}_t$ expressed as the noise-corrupted scaled Chebyshev matrix polynomials[2], i.e., $p_t(\mathbf{A})\widehat{\mathbf{X}}_0 + \sum_{j=1}^{t-1} q_j(\mathbf{A})\widehat{\boldsymbol{\xi}}_{t-j-1}$, where $p_t(x)$ is the scaled Chebyshev polynomials of the first kind with $q_1(x) = \frac{x}{2}$ and $q_t(x)$ has the same three-term recurrence as $p_t(x)$ but with a different second polynomial, i.e., $q_1(x) = x$. This way, i.e., handling cumulative noises instead of per-iteration noises in previous studies, eventually gives us a new form of the noise norm condition. The third key step is to extend the potential function in Gu (2015) to work under both the momentum setting and the noise setting in order to figure out noise tolerance bounds and convergence rate of $\mathbf{X}_t$.

Our analysis is much simpler and more general while achieving the optimal iteration complexity without the extra logarithmic factor. We further relate our new form of the noise norm conditions to the existing trigonometric one, which enables an improved analysis of the generalized eigenspace computation and canonical correlation analysis (CCA). Interestingly, the results show that the dependence of the noise tolerance bounds on the non-consecutive spectral gap varies with iterations from the linear dependency at the beginning to the square-root dependency near the convergence, but overall it remains commensurate with the existing results in terms of the noise tolerance (see Remark 1). This is the first time to capture the spectral gap dependence dynamics in theory, to the best of our knowledge. Since there have been no experimental studies under $p > k$ in the noise setting so far, we conduct an extensive experimental study to showcase the great performance of the considered algorithm under $p > k$ across applications. To summarize, we make the following contributions in this work:

- We present a simple yet general and optimal analysis that achieves the square-root dependence of the convergence rate on the non-consecutive gap and needs less noise conditions, that are in a new form while maintaining commensurate tolerance, for the accelerated noise-tolerant power method.
- We extend our analysis to the generalized eigenspace computation and CCA and achieve improved results of similar type.
- We conduct an extensive experimental study to showcase the great performance of the accelerated noise-tolerant power method under $p > k$ across different applications.

The rest of the paper is organized as follows. Section 2 further differentiates our work from existing studies. We then present our main results and their proofs in Section 3, and extend to two important applications in Section 4. Experiments are reported in Section 5 after which the paper is concluded in Section 6.

2 RELATED WORK

In the noiseless case, Gu (2015) theoretically justified the use of a larger iteration rank (i.e., $p > k$) by showing under mild conditions that the convergence depends on $(\lambda_k - \lambda_{q+1})$ for some $k \leq q \leq p$ that could be significantly larger than $(\lambda_k - \lambda_{k+1})$. Their potential function is

[2] To understand this type of polynomials, let's start from the power method without considering noises and orthonormalization for intuition, which can be written as $\hat{\mathbf{X}}_{t+1} = \mathbf{A}\hat{\mathbf{X}}_t$. By induction, we have $\hat{\mathbf{X}}_t = \mathbf{A}^t\hat{\mathbf{X}}_0$, where only monomials $\mathbf{A}^t$ are present. Replacing monomials by the scaled Chebyshev polynomials $p_t(\mathbf{A})$ gives us the momentum accelerated power method in the same vein, i.e., $\hat{\mathbf{X}}_{t+1} = p_t(\mathbf{A})\hat{\mathbf{X}}_t$. Taking three-term recurrence $p_{t+1}(\mathbf{A}) = \mathbf{A}p_t(\mathbf{A}) - \beta p_{t-1}(\mathbf{A})$, satisfied by $p_t(\mathbf{A})$, and noises in each iteration into account, we have that $\hat{\mathbf{X}}_{t+1} = \mathbf{A}\hat{\mathbf{X}}_t - \beta\hat{\mathbf{X}}_{t-1} + \hat{\boldsymbol{\xi}}_t$, which gives rise to the noise corrupted Chebyshev polynomials. We can get back to its polymonial expression by induction (see Lemma 3.5).

Table 1: Comparison with existing results.

	T	Cond 1: $\|\boldsymbol{\xi}_t\|$	Cond 2: $\|\mathbf{U}_q^\top \boldsymbol{\xi}_t\|$	Cond 3*
Balcan et al. (2016)	$O\left(\frac{\lambda_k}{\Delta_{k,q+1}} \log \frac{1}{\epsilon}\right)$	$O\left(\Delta_{k,q+1}\epsilon\right)$	$O\left(\Delta_{k,q+1}\epsilon\right)$	No
Xu and Li (2022)	$O\left(\sqrt{\frac{\lambda_k}{\Delta_{k,q+1}}} \log \frac{1}{\Delta_{q,q+1}\epsilon}\right)$	$O\left(\Delta_{k,q+1} \sin\tilde{\theta}_t\right)$	$O\left(\Delta_{k,q+1} \cos\tilde{\theta}_t\right)^{**}$	Yes
This work***	$\Theta\left(\sqrt{\frac{\lambda_k}{\Delta_{k,q+1}}} \log \frac{1}{\epsilon}\right)$	$O\left(\frac{\sqrt{\Delta_{k,q+1}}}{T-t+1}\left(\frac{\sqrt{\beta}}{\lambda_1^+}\right)^t \sqrt{\beta}\right)$	$O\left(\frac{\sqrt{\Delta_{k,q+1}}}{T-t+1}\left(\frac{\lambda_q^+}{\lambda_1^+}\right)^t \lambda_q^+\right)$	No

* refers to a restrictive noise condition expressed not in norm; ** $\tilde{\theta}_t$ is the largest principal angle induced by the augmented anti-triangular matrix at the t-th iteration; ***when $2\sqrt{\beta}$ is close to λ_{q+1};

good for handling the last iterate directly, which amounts to using monomials for analysis. By a different analysis that aims to handle per-iteration noises, Balcan et al. (2016) achieved such results for both convergence and noise tolerance, where the noise norm conditions are ϵ-dependent. However, the initial noises are not necessarily ϵ-small. Thus, both Xu and Li (2022) and our work consider iteration-dependent noise conditions. Xu and Li (2022) aimed to analyze per-iteration noises under the momentum acceleration, which however depends on the Schur decomposition of an augmented anti-triangular block matrix and gives rise to an extra logarithmic factor on an intermediate consecutive spectral gap in the iteration complexity. Also, their analysis requires a third noise condition which is quite restrictive as it is not in the simple form of noise norm like the other two, and empirically remains uncorroborated. In contrast, our analysis has no such limitations due to the focus on the last iterate and thus the cumulative noises via the noise-corrupted Chebyshev polynomial, and is supported well by our experimental study.

Xu et al. (2018) proposed scaled Chebyshev polynomials for acceleration under $p = k \geq 1$ for the noiseless case or under $p = k = 1$ for a special setting of stochastic noises. As noted in Balcan et al. (2016), stochastic analyses are orthogonal to and indeed cannot account for the noise models considered in Hardt and Price (2014); Balcan et al. (2016) as well as our work. Further, to the best of our knowledge, it has been unknown if their analysis can be applied to the setting of $k > 1$, not to mention the setting of $p > k$. Generally, the analysis from $k = 1$ to $k > 1$ needs significant and nontrivial changes for the type of the considered problem. Table 1 gives a succinct comparison, where notations are given in Section 1 and 3.

3 ANALYSIS

Algorithm 1 gives the pseudo code of the accelerated noise-tolerant power method. In this section, we present our main results and proofs for Algorithm 1 which is given a positive semi-definite matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$.

Algorithm 1 ANPM

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 spanned by the first k columns of $\mathbf{X}_T$.
3: QR factorize an entry-wise i.i.d. standard Gaussian matrix $\mathbf{G} \in \mathbb{R}^{n\times p}$ such that $\mathbf{X}_0\mathbf{R}_0 = \mathbf{G}$
4: QR factorize $\mathbf{Y} = \frac{1}{2}\mathbf{A}\mathbf{X}_0 + \boldsymbol{\xi}_0$ such that $\mathbf{X}_1\mathbf{R}_1 = \mathbf{Y}$
5: **for** $t = 1, \cdots, T-1$ **do**
6: $\quad \mathbf{Y} = \mathbf{A}\mathbf{X}_t - \beta\mathbf{X}_{t-1}\mathbf{R}_t^{-1} + \boldsymbol{\xi}_t$ for some noise $\boldsymbol{\xi}_t$
7: $\quad$ QR factorize $\mathbf{Y}$ such that $\mathbf{X}_{t+1}\mathbf{R}_{t+1} = \mathbf{Y}$
8: **end for**

All the missing proofs can be found in Appendix. Before presenting our main results, let's introduce necessary notations. $\boldsymbol{\Lambda}_j = \mathrm{diag}(\lambda_1, \cdots, \lambda_j), \boldsymbol{\Lambda}_{-j} = \mathrm{diag}(\lambda_{j+1}, \cdots, \lambda_n), \mathbf{U}_j = \left[\mathbf{u}_1, \cdots, \mathbf{u}_j\right]$, and $\mathbf{U}_{-j} = \left[\mathbf{u}_{j+1}, \cdots, \mathbf{u}_n\right]$, where $\mathbf{u}_j$ denotes $\mathbf{A}$'s eigenvector of unit length corresponding to the j-th largest eigenvalue. $\Delta_{i,j} = \lambda_i - \lambda_j$, and $\theta(\cdot,\cdot)$ represents the largest principal angle between two subspaces (Golub and Van Loan, 2013). We use matrix 2-norm throughout the paper. Specifically, we seek for a top-k eigenspace of the given matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ with an augmented matrix iterate $\mathbf{X}_t \in \mathbb{R}^{n\times p}$ ($p \geq k$), where k and p are referred to as the target rank and iteration rank, respectively.

3.1 Main Results

Theorem 3.1 *Let $k \leq q \leq p$ and assume that $\lambda_q > 2\sqrt{\beta} \geq \lambda_{q+1}$ for $\mathbf{A} \succcurlyeq \mathbf{0} \in \mathbb{R}^{n\times n}$. If the noise matrix $\boldsymbol{\xi}_t \in \mathbb{R}^{n\times p}$ satisfies that*

$$\begin{cases} \|\boldsymbol{\xi}_t\| = O\left(\frac{1}{(T-t+1)T}\left(\frac{\sqrt{\beta}}{\lambda_1^+}\right)^t \sqrt{\beta}\sin\theta_0\right) \\ \|\mathbf{U}_q^\top\boldsymbol{\xi}_t\| = O\left(\frac{1}{(T-t+1)T}\left(\frac{\lambda_q^+}{\lambda_1^+}\right)^t \lambda_q^+\cos\theta_0\right) \end{cases},$$

where $\lambda_j^+ = \frac{\lambda_j+\sqrt{\lambda_j^2-4\beta}}{2}$ *and* $\theta_t = \theta(\mathbf{X}_t, \mathbf{U}_q)$*, then after Algorithm 1 runs for* $T = \Theta\left(\sqrt{\frac{\lambda_k}{\lambda_k-2\sqrt{\beta}}}\log\frac{\tan\theta_0}{\epsilon}\right)$ *iterations*[3]*, we have that* $\sin\theta(\mathbf{X}_T, \mathbf{U}_k) < \epsilon$.

It is worth mentioning that the noise conditions in the above theorem are so general that they admit a wide range of noise types including but not limited to those aforementioned as long as their noise norms satisfy the conditions.

Remark 1 First, our iteration complexity above removes the extra logarithmic factor $\log\frac{1}{\Delta_{q,q+1}}$ in Xu and Li (2022). When $2\sqrt{\beta}$ is close to λ_{q+1}, $T = O\left(\sqrt{\frac{\lambda_k}{\Delta_{k,q+1}}}\log\frac{1}{\epsilon}\right)$. Second, we don't need the third restrictive noise condition required in Xu and Li (2022). Let's now look at the other two noise norm conditions in a new form here. On one hand, one may think that the exponential factors decay too fast compared to trigonometric ones. However, Lemma 3.2 below indicates that the two forms are in fact on the same order roughly, though exponential factors are on the lower side. On the other hand, interestingly, as shown in Table 1 the dependence of our noise tolerance bounds on the non-consecutive spectral gap varies with iterations from the linear dependency at the beginning, i.e., $\Delta_{k,q+1}$, to the square-root dependency near the convergence, i.e., $\sqrt{\Delta_{k,q+1}}$ which could be much larger than $\Delta_{k,q+1}$ as it is small for real data (Musco and Musco, 2015). We haven't seen such dynamic dependence of the noise tolerance on the spectral gap before. In this sense, overall the noise norm conditions in one form may not dominate those in the other form, or put another way, they largely remain commensurate in terms of overall tolerance.

Lemma 3.2

$$\begin{cases} \sin\theta_t = \Omega\left(\left(\frac{\sqrt{\beta}}{\lambda_1^+}\right)^t \sin\theta_0\right) \\ \cos\theta_t = \Omega\left(\left(\frac{\lambda_q^+}{\lambda_1^+}\right)^t \cos\theta_0\right) \end{cases}.$$

Remark 2 One may worry about the momentum parameter β whose optimal value $\frac{\lambda_{q+1}^2}{2}$ is not known. This is a common issue with momentum acceleration. In practice, a varying β can be used by setting $2\sqrt{\beta_t}$ to be the q-th largest diagonal entry of $\widehat{\mathbf{\Lambda}}_p = \mathbf{X}_t^\top(\mathbf{A}\mathbf{X}_t + \boldsymbol{\xi}_t) \in \mathbb{R}^{p\times p}$ and $2\sqrt{\beta_t} < \lambda_q$ will always approximately hold. Theoretically, our analysis and results for other values of β, given in Section C of Appendix, indicate that even β satisfying $2\sqrt{\beta} \le \lambda_{q+1}$ converges faster than the un-accelerated counterpart. Empirically, our experimental study shows that such setting of β works quite well.

[3]For brevity, $\tan\theta_0$ is not further bounded using $\mathbf{G}$ in Eq. (2). By Lemma 2.5 in Hardt and Price (2014), $\tan\theta(\mathbf{X}_0, \mathbf{U}_q) \le \frac{\iota\sqrt{n}}{\sqrt{p}-\sqrt{q-1}}$ with probability at least $1 - \iota^{-\Omega(p+1-q)} - e^{-\Omega(n)}$.

Remark 3 For the target top-k eigenspace of $\mathbf{A}$, we only need to take the first k columns of $\mathbf{X}_T$ after convergence (see Line 2 of Algorithm 1), because the subspace spanned by these columns of $\mathbf{X}_t$ (taking the first k columns in both sides of Eq. (6) and noting that $\mathbf{C}_t^{-1}$ in Eq. (5) is upper triangular) will approach the space spanned by the top-k eigenvectors. In fact, we can repeat the following proof only with the first k columns of $\mathbf{X}_t$ along a much simpler path as the target rank and iteration rank are equal in this case, and will have convergence of the first k columns of $\mathbf{X}_t$ to the target top-k eigenspace of $\mathbf{A}$.

3.2 Proof of Theorem 3.1

The road map of the proof is that we will derive the closed-form expression for the unnormalized iterate $\widehat{\mathbf{X}}_t$ which isolates signal from cumulative noises, and then plug it into the potential function for the last iterate so that we can find out how much total noises can be tolerated without affecting the linear convergence. We will also show that the upper bound on the convergence rate is actually tight.

We start from establishing the relationship between update Eq. (1) and its unnormalized counterpart Eq. (3). Let $\mathbf{C}_t = \prod_{j=t}^1 \mathbf{R}_j$ for $t \ge 1$, $\mathbf{C}_0 = \mathbf{I}$, $\widehat{\mathbf{X}}_0 = \mathbf{X}_0$, and define

$$\boldsymbol{\xi}_t = \widehat{\boldsymbol{\xi}}_t \mathbf{C}_t^{-1}, \tag{5}$$

for $t \ge 0$. We then have the following lemma about the connection between $\mathbf{X}_t$ and $\widehat{\mathbf{X}}_t$.

Lemma 3.3 $\widehat{\mathbf{X}}_t = \mathbf{X}_t\mathbf{C}_t$ *holds for* $t \ge 0$.

Two sequences of scaled Chebysev polynomials of the first and second kinds, each defined as follows by a three-term recurrence with initial polynomials:

$$p_{t+1}(x) = xp_t(x) - \beta p_{t-1}(x), \quad p_0(x) = 1,\ p_1(x) = \frac{1}{2}x,$$
$$q_{t+1}(x) = xq_t(x) - \beta q_{t-1}(x), \quad q_0(x) = 1,\ q_1(x) = x,$$

where they differ only in their second polynomials, have the following closed-form expressions.

Lemma 3.4 $\begin{cases} p_t(x) = \mathbf{z}_t \begin{bmatrix} \frac{1}{2}x \\ 1 \end{bmatrix} = \frac{(x^+)^t+(x^-)^t}{2} \\ q_t(x) = \mathbf{z}_t \begin{bmatrix} x \\ 1 \end{bmatrix} = \sum_{j=0}^t (x^+)^{t-j}(x^-)^j \end{cases}$,

where $\mathbf{z}_t = \begin{bmatrix} 0 & 1 \end{bmatrix}\begin{bmatrix} x & -\beta \\ 1 & 0 \end{bmatrix}^t$, *and* $x^\pm = \frac{x\pm\sqrt{x^2-4\beta}}{2}$ *which is a conjugate pair when* $|x| < 2\sqrt{\beta}$.

Define matrix polynomial

$$p_t(\mathbf{A}) = \textstyle\sum_{j=1}^n p_t(\lambda_j)\mathbf{u}_j\mathbf{u}_j^\top = \mathbf{U}_n p_t(\mathbf{\Lambda}_n)\mathbf{U}_n^\top,$$

where $p_t(\mathbf{\Lambda}_n) = \mathrm{diag}(p_t(\lambda_1), \cdots, p_t(\lambda_n))$. Thus,

$$p_{t+1}(\mathbf{A}) = \mathbf{A}p_t(\mathbf{A}) - \beta p_{t-1}(\mathbf{A}), p_0(\mathbf{A}) = \mathbf{I}, p_1(\mathbf{A}) = \frac{\mathbf{A}}{2},$$
$$q_{t+1}(\mathbf{A}) = \mathbf{A}q_t(\mathbf{A}) - \beta q_{t-1}(\mathbf{A}), q_0(\mathbf{A}) = \mathbf{I}, q_1(\mathbf{A}) = \mathbf{A}.$$

We then have the following lemma about $\widehat{\mathbf{X}}_t$'s closed-form expression.

Lemma 3.5 $\widehat{\mathbf{X}}_t = p_t(\mathbf{A})\widehat{\mathbf{X}}_0 + \sum_{j=0}^{t-1} q_j(\mathbf{A})\widehat{\boldsymbol{\xi}}_{t-j-1}$.

By Lemma 3.3 and 3.5, we can express $\mathbf{X}_t$ in a closed form:

$$\mathbf{X}_t = \left(p_t(\mathbf{A})\mathbf{X}_0 + \sum_{j=0}^{t-1} q_j(\mathbf{A})\widehat{\boldsymbol{\xi}}_{t-j-1}\right)\mathbf{C}_t^{-1}. \quad (6)$$

To analyze the convergence rate of $\mathbf{X}_t$, we extend the potential function for the noiseless power method under $p \geq k \geq 1$ in Gu (2015), i.e.,

$$\tilde{h}_T = \left\| \boldsymbol{\Lambda}_{-q}^T (\mathbf{U}_{-q}^\top \mathbf{X}_0)(\mathbf{U}_q^\top \mathbf{X}_0)^\dagger \begin{bmatrix} \boldsymbol{\Lambda}_k^{-1} \\ \mathbf{0} \end{bmatrix} \right\|,$$

where † represents the pseudo inverse of a matrix, to work for simultaneous momentum acceleration and noise corruption in our setting, and get our potential function

$$\begin{aligned} h_T &= \left\| (\mathbf{U}_{-q}^\top \mathbf{X}_T)(\mathbf{U}_q^\top \mathbf{X}_T)^\dagger \begin{bmatrix} \mathbf{I}_k \\ \mathbf{0} \end{bmatrix} \right\| \\ &= \left\| (\mathbf{U}_{-q}^\top \widehat{\mathbf{X}}_T)(\mathbf{U}_q^\top \widehat{\mathbf{X}}_T)^\dagger \begin{bmatrix} \mathbf{I}_k \\ \mathbf{0} \end{bmatrix} \right\|, \end{aligned}$$

where the second equality has used Lemma 3.3 and that $(\mathbf{U}_q^\top \mathbf{X}_T)^\dagger = \mathbf{C}_T (\mathbf{U}_q^\top \widehat{\mathbf{X}}_T)^\dagger \in \mathbb{R}^{p\times q}$. It is easy to see that $h_T = \tilde{h}_T$ if $\boldsymbol{\xi}_t \equiv \mathbf{0}$ and $\beta = 0$ since Eq. (1) recovers the noiseless power method then. We now can expand the two parts inside the norm in the following lemma.

Lemma 3.6 *Let $\mathbf{U}_q^\top \mathbf{X}_0 = \mathbf{P}\boldsymbol{\Sigma}\mathbf{Q}^\top$ be the compact SVD of $\mathbf{U}_q^\top \mathbf{X}_0$, where $\mathbf{P} \in \mathbb{R}^{q\times q}$ is orthogonal, $\boldsymbol{\Sigma} \in \mathbb{R}^{q\times q}$ is diagonal, and $\mathbf{Q} \in \mathbb{R}^{p\times q}$ is column-orthonormal. It holds that*

$$\mathbf{U}_{-q}^\top \widehat{\mathbf{X}}_T = p_T(\boldsymbol{\Lambda}_{-q})\mathbf{U}_{-q}^\top \mathbf{X}_0 + \boldsymbol{\Gamma},$$

and $(\mathbf{U}_q^\top \widehat{\mathbf{X}}_T)^\dagger \begin{bmatrix} \mathbf{I}_k \\ \mathbf{0} \end{bmatrix} =$

$$\left(\mathbf{I} + (\mathbf{U}_q^\top \mathbf{X}_0)^\dagger \boldsymbol{\Omega}\right)^\top \mathbf{Q}\mathbf{P}^\top \Big(\mathbf{I} + 2\mathrm{sym}(\mathbf{P}\mathbf{Q}^\top \boldsymbol{\Omega}^\top \mathbf{P}\boldsymbol{\Sigma}^{-1}\mathbf{P}^\top) + \mathbf{P}\boldsymbol{\Sigma}^{-1}\mathbf{P}^\top \boldsymbol{\Omega}\boldsymbol{\Omega}^\top \mathbf{P}\boldsymbol{\Sigma}^{-1}\mathbf{P}^\top\Big)^{-1} \mathbf{P}\boldsymbol{\Sigma}^{-1}\mathbf{P}^\top \begin{bmatrix} p_T^{-1}(\boldsymbol{\Lambda}_k) \\ \mathbf{0} \end{bmatrix},$$

where $\mathrm{sym}(\cdot)$ *extracts the symmetric part of a matrix and*

$$\begin{aligned} \boldsymbol{\Gamma} &= \sum_{t=0}^{T-1} q_t(\boldsymbol{\Lambda}_{-q})\mathbf{U}_{-q}^\top \widehat{\boldsymbol{\xi}}_{T-t-1}, \\ \boldsymbol{\Omega} &= p_T^{-1}(\boldsymbol{\Lambda}_q) \sum_{t=0}^{T-1} q_t(\boldsymbol{\Lambda}_q)\mathbf{U}_q^\top \widehat{\boldsymbol{\xi}}_{T-t-1}, \end{aligned}$$

represent the cumulative noises in the two parts inside the norm of the potential function, respectively.

By Lemma 3.6, we can write that

$$\begin{aligned} h_T = \Big\| &\Big(p_T(\boldsymbol{\Lambda}_{-q})\mathbf{U}_{-q}^\top \mathbf{X}_0 + \boldsymbol{\Gamma}\Big)\Big(\mathbf{I} + (\mathbf{U}_q^\top \mathbf{X}_0)^\dagger \boldsymbol{\Omega}\Big)^\top \\ &\mathbf{Q}\mathbf{P}^\top \Big(\mathbf{I} + 2\mathrm{sym}(\mathbf{P}\mathbf{Q}^\top \boldsymbol{\Omega}^\top \mathbf{P}\boldsymbol{\Sigma}^{-1}\mathbf{P}^\top) \\ &+ \mathbf{P}\boldsymbol{\Sigma}^{-1}\mathbf{P}^\top \boldsymbol{\Omega}\boldsymbol{\Omega}^\top \mathbf{P}\boldsymbol{\Sigma}^{-1}\mathbf{P}^\top\Big)^{-1} \\ &\mathbf{P}\boldsymbol{\Sigma}^{-1}\mathbf{P}^\top \begin{bmatrix} p_T^{-1}(\boldsymbol{\Lambda}_k) \\ \mathbf{0} \end{bmatrix} \Big\| \\ \leq \|p_T^{-1}(\boldsymbol{\Lambda}_k)\| &\Big(\|p_T(\boldsymbol{\Lambda}_{-q})\|\|\mathbf{U}_{-q}^\top \mathbf{X}_0\| + \|\boldsymbol{\Gamma}\|\Big)\|\boldsymbol{\Sigma}^{-1}\| \\ &\Big(1 - 2\|\boldsymbol{\Sigma}^{-1}\|\|\boldsymbol{\Omega}\| - (\|\boldsymbol{\Sigma}^{-1}\|\|\boldsymbol{\Omega}\|)^2\Big)^{-1} \\ &\Big(1 + \|(\mathbf{U}_q^\top \mathbf{X}_0)^\dagger\|\|\boldsymbol{\Omega}\|\Big). \end{aligned} \quad (7)$$

To further bound h_T, we need the following lemma.

Lemma 3.7

$$\begin{aligned} &\|p_T^{-1}(\boldsymbol{\Lambda}_k)\| \leq 2(\lambda_k^+)^{-T}, \quad \|p_T(\boldsymbol{\Lambda}_{-q})\| \leq (\sqrt{\beta})^T, \\ &\|q_t(\boldsymbol{\Lambda}_{-q})\| \leq (t+1)(\sqrt{\beta})^t, \\ &\|p_T^{-1}(\boldsymbol{\Lambda}_q) q_t(\boldsymbol{\Lambda}_q)\| \leq 2(t+1)(\lambda_q^+)^{t-T}. \end{aligned}$$

Assume that

$$\begin{cases} \|\widehat{\boldsymbol{\xi}}_t\| = O\left(\frac{1}{(T-t+1)T}\left(\sqrt{\beta}\right)^{t+1}\sin\theta_0\right), \\ \|\mathbf{U}_q^\top \widehat{\boldsymbol{\xi}}_t\| = O\left(\frac{1}{(T-t+1)T}\left(\lambda_q^+\right)^{t+1}\cos\theta_0\right). \end{cases} \quad (8)$$

By Lemma 3.7 and the above assumption, we can bound $\boldsymbol{\Gamma}$ and $\boldsymbol{\Omega}$ in Lemma 3.6 as follows:

$$\begin{aligned} \|\boldsymbol{\Gamma}\| &\leq \textstyle\sum_{t=0}^{T-1} \|q_t(\boldsymbol{\Lambda}_{-q})\|\|\widehat{\boldsymbol{\xi}}_{T-t-1}\| \\ &\leq \textstyle\sum_{t=0}^{T-1} (t+1)(\sqrt{\beta})^t \frac{(\sqrt{\beta})^{T-t}\sin\theta_0}{(t+2)T} \\ &\leq \beta^{\frac{T}{2}}\sin\theta_0, \end{aligned} \quad (9)$$

$$\begin{aligned} \|\boldsymbol{\Omega}\| &\leq \textstyle\sum_{t=0}^{T-1} \|p_T^{-1}(\boldsymbol{\Lambda}_q) q_t(\boldsymbol{\Lambda}_q)\|\|\mathbf{U}_q^\top \widehat{\boldsymbol{\xi}}_{T-t-1}\| \\ &\leq 2\textstyle\sum_{t=0}^{T-1} (t+1)(\lambda_q^+)^{t-T} \frac{(\lambda_q^+)^{T-t}\cos\theta_0}{16(t+2)T} \\ &\leq \frac{1}{8}\cos\theta_0. \end{aligned} \quad (10)$$

Also note that

$$\|\mathbf{U}_{-q}^\top \mathbf{X}_0\| = \sin\theta_0, \quad \|(\mathbf{U}_q^\top \mathbf{X}_0)^\dagger\| = \cos^{-1}\theta_0. \quad (11)$$

By Lemma 3.7 and Eq. (9)-(11), h_T in Eq. (7) can be further bounded as follows:

$$h_T \leq 16\left(\frac{\sqrt{\beta}}{\lambda_k^+}\right)^T \tan\theta_0 \triangleq \rho(T).$$

When $T > \frac{\log\left(\frac{16\tan\theta_0}{\epsilon}\right)}{\log\left(\frac{\lambda_k^+}{\sqrt{\beta}}\right)}$, we have $h_T \le \rho(T) < \epsilon$. Since

$$\log\left(\frac{\lambda_k^+}{\sqrt{\beta}}\right) \ge \frac{-1+\frac{\lambda_k^+}{\sqrt{\beta}}}{\frac{\lambda_k^+}{\sqrt{\beta}}} \ge \frac{1}{2}\sqrt{\frac{\lambda_k - 2\sqrt{\beta}}{\lambda_k}},$$

we get $h_T < \epsilon$ if $T = O\left(\sqrt{\frac{\lambda_k}{\lambda_k - 2\sqrt{\beta}}}\log\frac{\tan\theta_0}{\epsilon}\right)$. By the second proof of Lemma 2.3 in Balcan et al. (2016),

$$\begin{aligned}
& \sin\theta(\mathbf{X}_T, \mathbf{U}_k) \\
= \; & \left\|(\mathbf{I} - \mathbf{X}_T\mathbf{X}_T^\top)\mathbf{U}_k\right\| \\
\le \; & \left\|\mathbf{U}_k - \mathbf{X}_T(\mathbf{U}_q^\top\mathbf{X}_T)^\dagger\begin{bmatrix}\mathbf{I}_k\\ \mathbf{0}\end{bmatrix}\right\| \\
= \; & \left\|\mathbf{U}_q^\top\left(\mathbf{U}_k - \mathbf{X}_T(\mathbf{U}_q^\top\mathbf{X}_T)^\dagger\begin{bmatrix}\mathbf{I}_k\\ \mathbf{0}\end{bmatrix}\right)\right\| \\
& + \left\|\mathbf{U}_{-q}^\top\left(\mathbf{U}_k - \mathbf{X}_T(\mathbf{U}_q^\top\mathbf{X}_T)^\dagger\begin{bmatrix}\mathbf{I}_k\\ \mathbf{0}\end{bmatrix}\right)\right\| \\
= \; & \left\|\mathbf{U}_{-q}^\top\mathbf{X}_T(\mathbf{U}_q^\top\mathbf{X}_T)^\dagger\begin{bmatrix}\mathbf{I}_k\\ \mathbf{0}\end{bmatrix}\right\| \\
= \; & h_T < \epsilon.
\end{aligned}$$

We now need to convert noise conditions on $\widehat{\boldsymbol{\xi}}_t$ assumed in Eq. (8) to those in terms of $\boldsymbol{\xi}_t$ with the following lemma.

Lemma 3.8 $\|\mathbf{C}_t\| = \Theta\left((\lambda_1^+)^t\right)$.

It is easy to see that when the noise conditions on $\boldsymbol{\xi}_t$ given in Theorem 3.1 are satisfied, Eq. (8) will hold because both $\|\widehat{\boldsymbol{\xi}}_t\| \le \|\boldsymbol{\xi}_t\|\|\mathbf{C}_t\|$ and $\|\mathbf{U}_{-q}^\top\widehat{\boldsymbol{\xi}}_t\| \le \|\mathbf{U}_{-q}^\top\boldsymbol{\xi}_t\|\|\mathbf{C}_t\|$ hold by Eq. (5).

The remaining proof of Theorem 3.1 is to show that the above iteration complexity is tight. To this end, it suffices to consider a special case where $\mathbf{A} \succcurlyeq \mathbf{0}$ has eigenvalues

$$\begin{aligned}
\lambda_1 = \cdots = & \lambda_k > \lambda_{k+1} \ge \\
& \cdots \ge \lambda_q \ge 2\sqrt{\beta} = \lambda_{q+1} = \cdots = \lambda_n \ge 0,
\end{aligned}$$

and $\boldsymbol{\xi}_t = \mathbf{0}$. In this case, the equality above Eq. (7) becomes

$$\begin{aligned}
h_T &= \left\|p_T(\boldsymbol{\Lambda}_{-q})\mathbf{U}_{-q}^\top\mathbf{X}_0(\mathbf{U}_q^\top\mathbf{X}_0)^\dagger\begin{bmatrix}p_T^{-1}(\boldsymbol{\Lambda}_k)\\ \mathbf{0}\end{bmatrix}\right\| \\
&\ge \|p_T(\boldsymbol{\Lambda}_{-q})\|\,\sigma_{\min}\left(\mathbf{U}_{-q}^\top\mathbf{X}_0(\mathbf{U}_q^\top\mathbf{X}_0)^\dagger\begin{bmatrix}p_T^{-1}(\boldsymbol{\Lambda}_k)\\ \mathbf{0}\end{bmatrix}\right) \\
&= \|p_T(\boldsymbol{\Lambda}_{-q})\|\sigma_{\min}\left(\mathbf{U}_{-q}^\top\mathbf{X}_0(\mathbf{U}_q^\top\mathbf{X}_0)^\dagger\begin{bmatrix}\mathbf{I}\\ \mathbf{0}\end{bmatrix}p_T^{-1}(\boldsymbol{\Lambda}_k)\right) \\
&\ge \|p_T(\boldsymbol{\Lambda}_{-q})\|\sigma_{\min}\left(p_T^{-1}(\boldsymbol{\Lambda}_k)\right) \\
&\quad \sigma_{\min}\left(\mathbf{U}_{-q}^\top\mathbf{X}_0(\mathbf{U}_q^\top\mathbf{X}_0)^\dagger\begin{bmatrix}\mathbf{I}\\ \mathbf{0}\end{bmatrix}\right) \\
&\ge \left(\frac{\sqrt{\beta}}{\lambda_k^+}\right)^T\sigma_{\min}\left(\mathbf{U}_{-q}^\top\mathbf{X}_0(\mathbf{U}_q^\top\mathbf{X}_0)^\dagger\begin{bmatrix}\mathbf{I}\\ \mathbf{0}\end{bmatrix}\right),
\end{aligned}$$

where $\sigma_{\min}(\cdot)$ represents the smallest singular value of a matrix and the exponential factor in the last inequality can similarly give the iteration complexity $\sqrt{\frac{\lambda_k}{\Delta_{k,q+1}}}\log\frac{1}{\epsilon}$. Thus, in order to have $h_T < \epsilon$, we need $\Omega\left(\sqrt{\frac{\lambda_k}{\Delta_{k,q+1}}}\log\frac{1}{\epsilon}\right)$ iterations. This shows that the above upper bound matches the lower bound.

4 APPLICATIONS

We now extend our general results in Section 3.1 to generalized eigenspace computation and CCA.

4.1 Genearlized Eigenspace Computation

Given that each generalized eigenvector $\mathbf{u}_j$ of a pair of real symmetric matrices $(\mathbf{A}, \mathbf{B})$ with $\mathbf{B}$ being positive definite is an eigenvector of the real symmetric matrix $\tilde{\mathbf{A}} = \mathbf{B}^{-1/2}\mathbf{A}\mathbf{B}^{-1/2}$ pre-multiplied by $\mathbf{B}^{-1/2}$ such that $\mathbf{u}_i^\top\mathbf{B}\mathbf{u}_j = \delta_{ij}$ ($\delta_{ij} = 1$ if $i = j$ and 0 otherwise), based on the accelerated noise-tolerant power method, we can first write our iterate in two forms by Eq. (1) and Eq .(6), respectively, as follows:

$$\begin{cases}
\mathbf{X}_{t+1}\mathbf{R}_{t+1} = \tilde{\mathbf{A}}\mathbf{X}_t - \beta\mathbf{X}_{t-1}\mathbf{R}_t^{-1} + \mathbf{B}^{\frac{1}{2}}\boldsymbol{\xi}_t \in \mathbb{R}^{n\times p} \\
\mathbf{X}_t = \left(p_t(\tilde{\mathbf{A}})\mathbf{X}_0 + \sum_{j=0}^{t-1} q_j(\tilde{\mathbf{A}})\mathbf{B}^{\frac{1}{2}}\widehat{\boldsymbol{\xi}}_{t-j-1}\right)\mathbf{C}_t^{-1},
\end{cases}$$

and then pre-multiply both sides of two equations above by $\mathbf{B}^{-\frac{1}{2}}$ to get that

$$\begin{cases}
\mathbf{B}^{-\frac{1}{2}}\mathbf{X}_{t+1}\mathbf{R}_{t+1} = \mathbf{B}^{-\frac{1}{2}}\tilde{\mathbf{A}}\mathbf{X}_t - \beta\mathbf{B}^{-\frac{1}{2}}\mathbf{X}_{t-1}\mathbf{R}_t^{-1} + \boldsymbol{\xi}_t \\
\mathbf{B}^{-\frac{1}{2}}\mathbf{X}_t = \left(\mathbf{B}^{-\frac{1}{2}}p_t(\tilde{\mathbf{A}})\mathbf{X}_0 + \sum_{j=0}^{t-1}\mathbf{B}^{-\frac{1}{2}}q_j(\tilde{\mathbf{A}})\mathbf{B}^{\frac{1}{2}}\widehat{\boldsymbol{\xi}}_{t-j-1}\right)\mathbf{C}_t^{-1}.
\end{cases}$$

Letting $\mathbf{Z}_t = \mathbf{B}^{-\frac{1}{2}}\mathbf{X}_t$, we can write that

$$\begin{cases}
\mathbf{Z}_{t+1}\mathbf{R}_{t+1} = \mathbf{B}^{-1}\mathbf{A}\mathbf{Z}_t - \beta\mathbf{Z}_{t-1}\mathbf{R}_t^{-1} + \boldsymbol{\xi}_t \\
\mathbf{Z}_t = \Big(\mathbf{B}^{-\frac{1}{2}}p_t(\tilde{\mathbf{A}})\mathbf{B}^{\frac{1}{2}}\mathbf{Z}_0 \\
\qquad\qquad + \sum_{j=0}^{t-1}\mathbf{B}^{-\frac{1}{2}}q_j(\tilde{\mathbf{A}})\mathbf{B}^{\frac{1}{2}}\widehat{\boldsymbol{\xi}}_{t-j-1}\Big)\mathbf{C}_t^{-1},
\end{cases} \tag{12}$$

where $\boldsymbol{\xi}_t$ is the noise term caused by approximating the term $\mathbf{B}^{-1}\mathbf{A}\mathbf{Z}_t$ by a warm-started least-squares solver as in Ge et al. (2016), and $\mathbf{R}_t$ now makes $\mathbf{Z}_t$ $\mathbf{B}$-orthonormal, i.e., $\mathbf{Z}_t^\top\mathbf{B}\mathbf{Z}_t = \mathbf{I}$, through, e.g., the modified Gram-Schmidt process with inner product $\langle , \rangle_{\mathbf{B}}$.

By the analysis of computing $\mathbf{B}^{-1}\mathbf{A}\mathbf{Z}_t$ in Ge et al. (2016), the ratio of final to initial error for the least-squares solver can be bounded by constants, where the final error can be expressed as $\|\boldsymbol{\xi}_t\|_{B,F}^2 = \boldsymbol{\xi}_t^\top\mathbf{B}\boldsymbol{\xi}_t$. However, the initial error

is $O\left(\lambda_1^2 \tan^2\theta(\mathbf{Z}_t, \mathbf{U}_q)\right)$. We can convert our noise conditions in Theorem 3.1 to be of similar trigonometric form by Lemma 3.2 so that Theorem 3.1 can be applied together with the approximation cost to get the total complexity of the generalized eigenspace computation in Theorem 4.1. Details can be found in Appendix where the corresponding algorithm is provided as well.

Theorem 4.1 *Let $k \leq q \leq p$ and assume that $\lambda_q > 2\sqrt{\beta} \geq \lambda_{q+1}$ and $\lambda_n \geq 0$ for a pair of $n \times n$ real symmetric matrices $(\mathbf{A}, \mathbf{B})$ with $\mathbf{B} \succ \mathbf{0}$. After the update in Eq. (12) runs for $T = O\left(\frac{1}{\sqrt{\rho}} \log \frac{\tan\theta_0}{\epsilon}\right)$ iterations, we have that $\sin\theta(\mathbf{Z}_T, \mathbf{U}_k) < \epsilon$ in time complexity*

$$O\Bigg(\mathrm{nnz}(\mathbf{B})p\sqrt{\frac{\kappa(\mathbf{B})}{\rho}}\Big(\log\frac{1}{\cos\theta_0}\log\frac{p\gamma}{\rho\cos\theta_0} + \log\frac{1}{\epsilon}\log\frac{p\gamma}{\rho}\Big) + \frac{\mathrm{nnz}(\mathbf{A})p + \mathrm{nnz}(\mathbf{B})p^2}{\sqrt{\rho}}\log\frac{1}{\epsilon\cos\theta_0}\Bigg),$$

where $\rho = \frac{\lambda_k - 2\sqrt{\beta}}{\lambda_k}$, $\gamma = \frac{\lambda_1}{\lambda_k}$, $\kappa(\mathbf{B})$ represents the condition number of $\mathbf{B}$, $\mathrm{nnz}(\cdot)$ represents the number of nonzero entries in a matrix, and $\theta_0 = \theta(\mathbf{Z}_0, \mathbf{U}_q)$.

We can see that when $2\sqrt{\beta}$ is close to λ_{q+1} the above result improves over Ge et al. (2016) in two ways, i.e., from small $\Delta_{k,k+1}$ to large $\Delta_{k,q+1}$ and from large $\frac{\sqrt{\kappa(\mathbf{B})}}{\Delta_{k,k+1}}$ to small $\sqrt{\frac{\kappa(\mathbf{B})}{\Delta_{k,q+1}}}$. Both ways together achieve an effect of double acceleration.

4.2 Canonical Correlation Analysis

CCA aims to find two k-dimensional canonical subspaces, one for each of datasets $\mathbf{X} \in \mathbb{R}^{d_x \times n}$ and $\mathbf{Y} \in \mathbb{R}^{d_y \times n}$ with a bit abuse of notation that $\mathbf{X}, \mathbf{Y}$ without subscripts represent input data here, such that data projections onto their respective subspaces are maximally correlated. It is a special case of generalized eigenspace computation with real symmetric matrix pair given in the following form:

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

where $\mathbf{C}_{xx} = \frac{1}{n}\mathbf{X}\mathbf{X}^\top + r_x\mathbf{I}$, $\mathbf{C}_{yy} = \frac{1}{n}\mathbf{Y}\mathbf{Y}^\top + r_y\mathbf{I}$ and $\mathbf{C}_{xy} = \frac{1}{n}\mathbf{X}\mathbf{Y}^\top$ are the two auto-covariance matrices and the cross-covariance matrix of $\mathbf{X}, \mathbf{Y}$, respectively, with $r > 0$ being regularization parameter for avoiding ill-conditioning. We adopt the update equations in Xu and Li (2021) for CCA but with change from $p = k$ to $p \geq k$ as follows:

$$\begin{cases} \mathbf{\Phi}_{t+1}\mathbf{R}_{t+1} = \mathbf{C}_{xx}^{-1}\mathbf{C}_{xy}\big(\mathbf{C}_{yy}^{-1}\mathbf{C}_{xy}^\top\mathbf{\Phi}_t + \boldsymbol{\xi}_t^1\big) \\ \qquad\qquad -\beta\mathbf{\Phi}_{t-1}\mathbf{R}_t^{-1} + \boldsymbol{\xi}_t^2 \in \mathbb{R}^{d_x \times p} \\ \mathbf{\Psi}_{t+1}\mathbf{S}_{t+1} = \mathbf{C}_{yy}^{-1}\mathbf{C}_{xy}^\top\big(\mathbf{C}_{xx}^{-1}\mathbf{C}_{xy}\mathbf{\Psi}_t + \boldsymbol{\eta}_t^1\big) \\ \qquad\qquad -\beta\mathbf{\Psi}_{t-1}\mathbf{S}_t^{-1} + \boldsymbol{\eta}_t^2 \in \mathbb{R}^{d_y \times p}, \end{cases} \tag{13}$$

which alternates between $\beta = 0$ and $\beta > 0$ while running the update in Eq. (12) and merges the two alternating steps into one step in terms of two iterates $\mathbf{\Phi}_t$ and $\mathbf{\Psi}_t$. For analysis, we need closed-form representations of both iterates $\mathbf{\Phi}_t$ and $\mathbf{\Psi}_t$ like that below Eq. (12), which can be written as:

$$\begin{cases} \mathbf{\Phi}_t = \Big(\mathbf{C}_{xx}^{-\frac{1}{2}}p_t(\mathbf{H}_{yy})\mathbf{C}_{xx}^{\frac{1}{2}}\mathbf{\Phi}_0 + \sum_{j=0}^{t-1}\mathbf{C}_{xx}^{-\frac{1}{2}}q_j(\mathbf{H}_{yy}) \\ \qquad \mathbf{C}_{xx}^{\frac{1}{2}}\big(\mathbf{C}_{xx}^{-1}\mathbf{C}_{xy}\widehat{\boldsymbol{\xi}}_{t-j-1}^1 + \widehat{\boldsymbol{\xi}}_{t-j-1}^2\big)\Big)\mathbf{C}_{\phi,t}^{-1} \\ \mathbf{\Psi}_t = \Big(\mathbf{C}_{yy}^{-\frac{1}{2}}p_t(\mathbf{H}_{xx})\mathbf{C}_{yy}^{\frac{1}{2}}\mathbf{\Psi}_0 + \sum_{j=0}^{t-1}\mathbf{C}_{yy}^{-\frac{1}{2}}q_j(\mathbf{H}_{xx}) \\ \qquad \mathbf{C}_{yy}^{\frac{1}{2}}\big(\mathbf{C}_{yy}^{-1}\mathbf{C}_{xy}^\top\widehat{\boldsymbol{\eta}}_{t-j-1}^1 + \widehat{\boldsymbol{\eta}}_{t-j-1}^2\big)\Big)\mathbf{C}_{\psi,t}^{-1}, \end{cases}$$

where

$$\begin{aligned} \mathbf{H}_{xx} &= \mathbf{C}_{yy}^{-\frac{1}{2}}\mathbf{C}_{xy}^\top\mathbf{C}_{xx}^{-1}\mathbf{C}_{xy}\mathbf{C}_{yy}^{-\frac{1}{2}}, \\ \mathbf{H}_{yy} &= \mathbf{C}_{xx}^{-\frac{1}{2}}\mathbf{C}_{xy}\mathbf{C}_{yy}^{-1}\mathbf{C}_{xy}^\top\mathbf{C}_{xx}^{-\frac{1}{2}}. \end{aligned}$$

Let the partial singular value decomposition of the whitened empirical cross-covariance matrix, defined as $\mathbf{C} = \mathbf{C}_{xx}^{-1/2}\mathbf{C}_{xy}\mathbf{C}_{yy}^{-1/2}$, be $\mathring{\mathbf{U}}_j\mathbf{\Sigma}_j\mathring{\mathbf{V}}_j^\top$ with $\mathring{\mathbf{U}}_j$ and $\mathring{\mathbf{V}}_j$ being $\mathbf{C}$'s top-j left and right singular subspaces, respectively, and $\mathbf{\Sigma}_j = \mathrm{diag}(\sigma_1, \cdots, \sigma_j)$ having $\mathbf{C}$'s top-j singular values on the diagonal. The solution to the k-CCA problem then can be written as $(\mathbf{U}_k, \mathbf{V}_k) = (\mathbf{C}_{xx}^{-\frac{1}{2}}\mathring{\mathbf{U}}_k, \mathbf{C}_{yy}^{-\frac{1}{2}}\mathring{\mathbf{V}}_k)$.

Theorem 4.2 *Let $k \leq q \leq p$ and assume that $\sigma_q^2 > 2\sqrt{\beta} \geq \sigma_{q+1}^2$. After the update in Eq. (13) runs for $T = O\left(\frac{1}{\sqrt{\rho}}\log\frac{1}{\epsilon\cos\theta_0}\right)$, we have that*

$$\sin\max\{\theta(\mathbf{\Phi}_T, \mathbf{U}_k), \theta(\mathbf{\Psi}_T, \mathbf{V}_k)\} \leq \epsilon$$

in time complexity

$$O\Bigg(\mathrm{nnz}(\mathbf{X}, \mathbf{Y})p\sqrt{\frac{\kappa(\mathbf{X}, \mathbf{Y})}{\rho}}\Big(\log\frac{1}{\cos\theta_0}\log\frac{p\gamma}{\rho\cos\theta_0} + \log\frac{1}{\epsilon}\log\frac{p\gamma}{\rho}\Big) + \frac{\mathrm{nnz}(\mathbf{X}, \mathbf{Y})p^2}{\sqrt{\rho}}\log\frac{1}{\epsilon\cos\theta_0}\Bigg),$$

where $\rho = \frac{\sigma_k^2 - 2\sqrt{\beta}}{\sigma_k^2}$, $\gamma = \frac{\sigma_1^2}{\sigma_k^2}$, and

$$\begin{aligned} \mathrm{nnz}(\mathbf{X}, \mathbf{Y}) &= \mathrm{nnz}(\mathbf{X}) + \mathrm{nnz}(\mathbf{Y}), \\ \kappa(\mathbf{X}, \mathbf{Y}) &= \max\left\{\kappa(\mathbf{C}_{xx}), \kappa(\mathbf{C}_{yy})\right\}, \\ \theta_0 &= \max\left\{\theta(\mathbf{\Phi}_0, \mathbf{U}_q), \theta(\mathbf{\Psi}_0, \mathbf{V}_q)\right\}. \end{aligned}$$

When $2\sqrt{\beta}$ is close to σ_{q+1}^2, Theorem 4.2 improves over Xu and Li (2021) in two ways, i.e., from small $(\sigma_k^2 - \sigma_{k+1}^2)$ to large $(\sigma_k^2 - \sigma_{q+1}^2)$ and having the additional factor $\log\frac{1}{\sigma_k^2 - \sigma_{k+1}^2}$ removed.

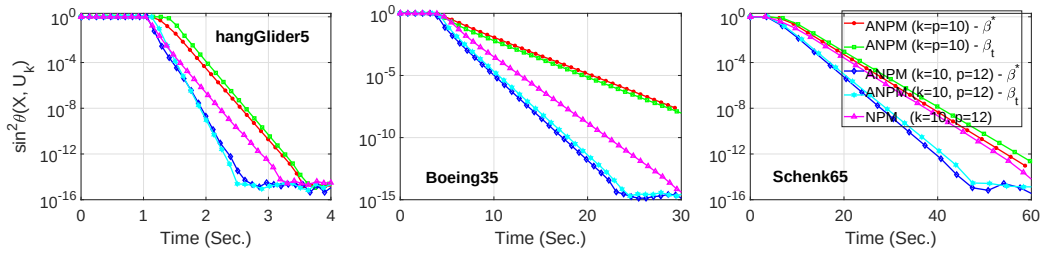

Figure 1: Performance of the ANPM under $p > k$ in comparison with the ANPM under $p = k$ for two settings of β and the NPM under $p > k$, where i.i.d. zero mean Gaussian noises of varying variance $\sigma_t = \frac{10^5}{1.1^t}$ were injected into iterations.

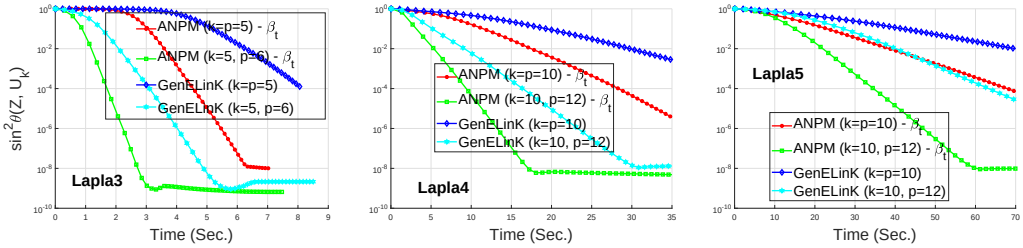

Figure 2: Performance of the ANPM for generalized eigenspace computation under $p > k$ in comparison with the ANPM under $p = k$, using dynamic momentum parameter β_t, and GenELink in the same two settings of target rank k and iteration rank p.

5 EXPERIMENTS

In this section, we conduct experiments to test the performance of Algorithm 1 under $p \geq k$ in different scenarios including the accelerated noise-tolerant power method, generalized eigenspace computation, and CCA. All algorithms were implemented in Matlab and fed with the same initial for each dataset in each setting. For benchmarking, the ground-truth information is obtained using matlab's eigs function for the first two scenarios and svds function for the last scenario. Target rank k and iteration rank p used are shown on the figure legends, and we set $q = p$ throughout experiments. More experimental results are provided in Section D of Appendix.

5.1 ANPM

We test the performance of Algorithm 1 using three real data matrices downloaded from the sparse matrix collection[4] with statistics given in Table 2. Two settings of parameter β were tested: optimal $2\sqrt{\beta^*} = \lambda_{q+1}$ and dynamic $2\sqrt{\beta_t}$ which is set according to Remark 2. We compare with the noisy power method (NPM). The results were evaluated with measure $\sin\theta(\mathbf{X}_t, \mathbf{U}_k)$ for which smaller is better. We experimented with i.i.d. zero mean Gaussian noises of varying variance $\sigma_t = \frac{10^5}{1.1^t}$ injected into iterations. We can see from Figure 1 that using a larger iteration rank improves significantly, especially together with momentum acceleration, and even the NPM under $p > k$ can outperform the ANPM under $p = k$. Interestingly, two settings of β perform almost equally well.

Table 2: Datasets for ANPM

Name	n	$\text{nnz}(\mathbf{A})$
hangGlider5	16011	162363
Boeing35	30237	1450163
Schenk65	48066	360428

Table 3: Datasets for generalized eigenspace computation.

Name	n	$\text{nnz}(\mathbf{A})$	$\text{nnz}(\mathbf{B})$
Lapla3	5795	136565	141779
Lapla4	10891	259425	269639
Lapla5	18903	455337	489875

5.2 Genearlized Eigenspace Computation

We compare the ANPM in different settings, including the advocated setting of $p > k$, with the GenELinK algorithm (Ge et al., 2016) for top-k generalized eigenspace computation (see Algorithm 2 in Appendix) on three datasets[5] given in Table 3, and use evaluation measure $\sin\theta(\mathbf{Z}_t, \mathbf{U}_k)$ accordingly. For approximating the multiplication with inverse matrix $\mathbf{B}^{-1}$, we use the built-in MATLAB function pcg (preconditioned conjugate gradients method) as the

[4] https://sparse.tamu.edu/

[5] http://faculty.smu.edu/yzhou/data/matrices.htm

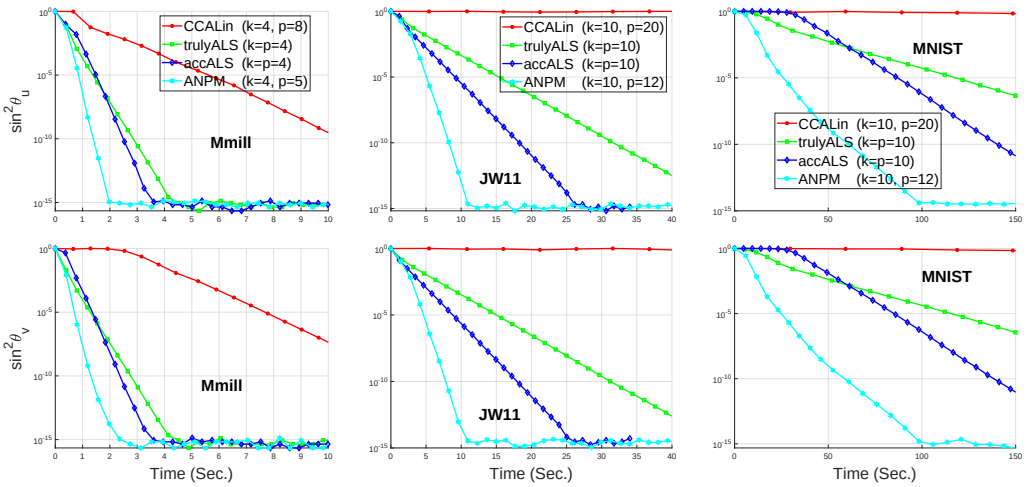

Figure 3: Performance of the ANPM for CCA under $p > k$ in comparison with three baselines, i.e., the trulyALS and accALS for $k = p$ and the CCALin which requires $p = 2k$.

least-square solver with 10 iterations for each run. Figure 2 shows their performance comparison, where we observed similar patterns to those observed in Figure 1 but with large performance gaps.

Table 4: Datasets for CCA.

Name	n	d_x	d_y
Mmill	3×10^4	100	120
JW11	3×10^4	273	112
MNIST	6×10^4	392	392

5.3 CCA

We use three common datasets, described in Table 4, for CCA (Ge et al., 2016; Wang et al., 2016; Arora et al., 2017; Xu and Li, 2019) with regularization parameters $r_x = r_y = 0.1$. We compare the ANPM under $p > k$ (see Algorithm 3 in Appendix) to recent CCA algorithms including CCALin (Ge et al., 2016) (which requires $p = 2k$), trulyALS under $p = k$ (Xu and Li, 2019), and the latest accALS under $p = k$ (Xu and Li, 2021). SVRG is used as the least-squares solver running 2 epochs for each CCA algorithm. Each epoch runs n iterations with constant step-sizes $\alpha_\phi = 1/\max_i \|\mathbf{x}_i\|_2^2$ for $\mathbf{\Phi}_t$ and $\alpha_\psi = 1/\max_i \|\mathbf{y}_i\|_2^2$ for $\mathbf{\Psi}_t$, where $\mathbf{x}_i$ represents $\mathbf{X}$'s i-th column. Two evaluation measures are $\sin\theta_u$ and $\sin\theta_v$, where $\theta_u = \theta(\mathbf{\Phi}_t, \mathbf{U}_k)$ and $\theta_v = \theta(\mathbf{\Psi}_t, \mathbf{V}_k)$. Figure 3 reports these algorithms' performance, which further confirms the advantage of the setting of $p > k$.

Before closing this section, it is worth mentioning that we also provided experiments about varying iteration rank p in Section D of Appendix which aim to investigate the trade-off between convergence speedup and additional computational cost brought by our approach in terms of running time. These experiments show that on our datasets using an iteration rank p that is about $k/2$ larger than target rank k often brings most significant performance gain, while further increasing p may gain not more but probably less because additional computational cost starts to offset a large part of the gain from less iterations due to faster convergence. In addition, convergence to the exact optima needs the noise conditions to be satisfied. Particularly, the noise needs to vanish eventually by our theory, otherwise the convergent point can only be within the noise ball around the exact optima. In practice, the noise can't completely vanish. The best case is that noise magnitude matches the machine precision, e.g., 10^{-16}, where we may consider convergence accuracy of 10^{-16} achieves convergence to the exact optima, as seen in a part of our experiments.

6 CONCLUSIONS

We present a general analysis for the accelerated noise-tolerant power method under a larger iteration rank than the target rank, which needs less noise conditions but can achieve the optimal iteration complexity. The noise tolerance is characterized by two norm conditions in a new form which we relate to the existing form. One interesting phenomenon in theory is the dynamic spectral gap dependence of the noise tolerance during iterations, varying from the linear at the beginning to the square-root near the convergence, while maintaining commensurate overall tolerance. The analysis is much simpler than previous ones, as it only needs to leverage the noise-corrupted scaled Chebyshev polynomials. Further, it enables us to give an improved analysis for generalized eigenspace computation and CCA. We demonstrate that the accelerated noise-tolerant power method with a larger iteration rank than the target rank performs best in practice across applications.

Acknowledgements

Authors would like to thank reviewers for their comments that helped improve the quality of the paper.

References

Raman Arora, Teodor Vanislavov Marinov, Poorya Mianjy, and Nati Srebro. Stochastic approximation for canonical correlation analysis. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 4778–4787, 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 2016, New York, USA, June 23-26, 2016*, pages 284–309, 2016.

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

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. doi: 10.1137/130938700.

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 Dan Boneh, Tim Roughgarden, and Joan Feigenbaum, editors, *Symposium on Theory of Computing Conference, STOC'13, Palo Alto, CA, USA, June 1-4, 2013*, pages 331–340. ACM, 2013. doi: 10.1145/2488608.2488650.

Yingyu Liang, Maria-Florina Balcan, Vandana Kanchanapally, and David P. Woodruff. Improved distributed principal component analysis. In Zoubin Ghahramani, Max Welling, Corinna Cortes, Neil D. Lawrence, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 3113–3121, 2014.

Ziqi Liu, Yu-Xiang Wang, and Alexander J. Smola. Fast differentially private matrix factorization. In Hannes Werthner, Markus Zanker, Jennifer Golbeck, and Giovanni Semeraro, editors, *Proceedings of the 9th ACM Conference on Recommender Systems, RecSys 2015, Vienna, Austria, September 16-20, 2015*, pages 171–178. ACM, 2015. doi: 10.1145/2792838.2800191.

Vien V. Mai and Mikael Johansson. Noisy accelerated power method for eigenproblems with applications. *IEEE Trans. Signal Processing*, 67(12):3287–3299, 2019. doi: 10.1109/TSP.2019.2908126.

Ioannis Mitliagkas, Constantine Caramanis, and Prateek Jain. Memory limited, streaming PCA. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States.*, pages 2886–2894, 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 28: 29th Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 1396–1404, 2015.

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

Peng Xu, Bryan D. He, Christopher De Sa, Ioannis Mitliagkas, and Christopher Ré. Accelerated stochastic power iteration. In *International Conference on Artificial Intelligence and Statistics, AISTATS 2018, 9-11 April 2018, Playa Blanca, Lanzarote, Canary Islands, Spain*, pages 58–67, 2018.

Zhiqiang Xu and Ping Li. Towards practical alternating least-squares for CCA. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pages 14737–14746, 2019.

Zhiqiang Xu and Ping Li. On the faster alternating least-squares for CCA. In Arindam Banerjee and Kenji Fukumizu, editors, *The 24th International Conference on Artificial Intelligence and Statistics, AISTATS 2021, April 13-15, 2021, Virtual Event*, volume 130 of *Proceedings of Machine Learning Research*, pages 1621–1629. PMLR, 2021.

Zhiqiang Xu and Ping Li. Faster noisy power method. In Sanjoy Dasgupta and Nika Haghtalab, editors, *International Conference on Algorithmic Learning Theory, 29-1 April 2022, Paris, France*, volume 167 of *Proceedings of*

Machine Learning Research, pages 1138–1164. PMLR, 2022.

A A WRONG ANALYSIS

The analysis of Mai and Johansson (2019) is wrong in proofs of their Theorems 1-2, where Theorem 1 is a special case of their Theorem 2. Let's focus on their Theorem 2 whose proof is given in Appendix C, where Eq. (24) is the key to their proof but does not hold actually. Let's quote the part of Mai and Johansson (2019) where the wrong analysis originates as follows:

> " Recall that $\mathbf{U}$ is the matrix of top-k eigenvectors of $\mathbf{B}^{-1}\mathbf{A}$ and $\bar{\mathbf{V}} = \begin{bmatrix}\mathbf{\Lambda}\mathbf{U}^\top & \mathbf{U}^\top\end{bmatrix}^\top$ is the top-k eigenvectors of the extended matrix $\mathbf{C} = \begin{bmatrix}\mathbf{B}^{-1}\mathbf{A} & -\beta\mathbf{I} \\ \mathbf{I} & \mathbf{0}\end{bmatrix}$ with $\mathbf{\Lambda} = \mathrm{diag}(\mu_1, \cdots, \mu_k)$ being the corresponding matrix of eigenvalues. Let
>
> $$\mathbf{V} = \mathcal{B}^{-1/2}\bar{\mathbf{V}}(\mathbf{I} + \mathbf{\Lambda}^2)^{-1/2} \text{ and } \mathcal{B} = \begin{bmatrix}\mathbf{B} & \mathbf{0} \\ \mathbf{0} & \mathbf{B}\end{bmatrix},$$
>
> then $\mathbf{V}^\top\mathcal{B}\mathbf{V} = \mathbf{I}_k$. If we further let $\mathbf{V}_\perp$ be an orthogonal basis w.r.t. $\mathcal{B}$ of the orthogonal complement of $\mathrm{span}(\mathbf{V})$, then one can decompose $\mathbf{C}$ as
>
> $$\mathbf{C} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^\top\mathcal{B} + \mathbf{V}_\perp\mathbf{\Lambda}_\perp\mathbf{V}_\perp^\top\mathcal{B}, \qquad (24)$$
>
> where $\mathbf{\Lambda}_\perp = \mathrm{diag}(\mu_{k+1}, \cdots, \mu_{2d-k})$."

Note that $\mathcal{B}^{-1/2}$ is missing in the definition of $\mathbf{V}$ given above their Eq.(24), which should be a typo otherwise $\mathbf{V}^\top\mathcal{B}\mathbf{V} \neq \mathbf{I}_k$.

To see why it is wrong, let's post-multiply both sides of Eq. (24) with $\mathcal{B}^{-1}$, and then we have

$$\mathbf{C}\mathcal{B}^{-1} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^\top + \mathbf{V}_\perp\mathbf{\Lambda}_\perp\mathbf{V}_\perp^\top.$$

We now can see that the right-hand side is a *symmetric* matrix but the left-hand side $\mathbf{C}\mathcal{B}^{-1} = \begin{bmatrix}\mathbf{B}^{-1}\mathbf{A}\mathbf{B}^{-1} & -\beta\mathbf{B}^{-1} \\ \mathbf{B}^{-1} & \mathbf{0}\end{bmatrix}$ is *asymmetric*. In addition, the right-hand side could be complex because $\mathbf{\Lambda}_\perp$ is complex (since $2\sqrt{\beta} > |\lambda_{k+1}|$ there) and transpose of $\mathbf{V}_\perp$ is used instead of conjugate transpose, but the left-hand side is always real. Thus, their following analysis is incorrect, due to the fact that this real asymmetric matrix is not a normal matrix and thus does not have a unitary diagonalization. In fact, their analysis directly follows Ge et al. (2016) where orthogonal diagonalization indeed holds

$$\mathbf{A} = \mathbf{B}\begin{bmatrix}\mathbf{U} & \mathbf{U}_\perp\end{bmatrix}\mathrm{diag}(\mathbf{\Lambda}, \mathbf{\Lambda}_\perp)\begin{bmatrix}\mathbf{U} & \mathbf{U}_\perp\end{bmatrix}^\top\mathbf{B}$$

because the relevant matrix $\mathbf{A}$ is real symmetric. But now when momentum is considered, relevant matrix $\mathbf{C}\mathcal{B}^{-1}$ becomes real asymmetric and not normal.

B MISSING PROOFS

In this section, we provide all the missing proofs in the main text. Particularly, we will restate Lemmas 3.3, 3.5. 3.6, 3.8 of the main paper in the setting of generalized eigenspace computation where a pair of real symmetric matrices $(\mathbf{A}, \mathbf{B})$ with $\mathbf{B}$ positive definite is considered, because it covers the standard case by setting $\mathbf{B} = \mathbf{I}$. In this setting, $\mathbf{u}_j$ denotes $(\mathbf{A}, \mathbf{B})$'s eigenvector of unit length corresponding to the j-th largest eigenvalue and satisfies $\mathbf{u}_i^\top\mathbf{B}\mathbf{u}_j = \delta_{ij}$, and

$$\mathbf{A} = \mathbf{B}\begin{bmatrix}\mathbf{U}_j & \mathbf{U}_{-j}\end{bmatrix}\mathrm{diag}(\mathbf{\Lambda}_j, \mathbf{\Lambda}_{-j})\begin{bmatrix}\mathbf{U}_j & \mathbf{U}_{-j}\end{bmatrix}^\top\mathbf{B}$$

for any $1 \leq j \leq n$. It is worth noting in this case that following notations in Section 4.1, Eqs. (1)-(4) become as follows

$$\mathbf{Z}_{t+1}\mathbf{R}_{t+1} = \mathbf{B}^{-1}\mathbf{A}\mathbf{Z}_t - \beta\mathbf{Z}_{t-1}\mathbf{R}_t^{-1} + \boldsymbol{\xi}_t \qquad (14)$$

$$\mathbf{Z}_0\mathbf{R}_0 = \mathbf{G}, \quad \mathbf{Z}_1\mathbf{R}_1 = \frac{1}{2}\mathbf{B}^{-1}\mathbf{A}\mathbf{Z}_0 + \boldsymbol{\xi}_0 \in \mathbb{R}^{n\times p}. \qquad (15)$$

$$\widehat{\mathbf{Z}}_{t+1} = \mathbf{B}^{-1}\mathbf{A}\widehat{\mathbf{Z}}_t - \beta\widehat{\mathbf{Z}}_{t-1} + \widehat{\boldsymbol{\xi}}_t \in \mathbb{R}^{n\times p}, \qquad (16)$$

$$\widehat{\mathbf{Z}}_0 = \mathbf{Z}_0, \quad \widehat{\mathbf{Z}}_1 = \frac{1}{2}\mathbf{B}^{-1}\mathbf{A}\widehat{\mathbf{Z}}_0 + \widehat{\boldsymbol{\xi}}_0, \qquad (17)$$

where $\mathbf{R}_t$ makes $\mathbf{Z}_t$ $\mathbf{B}$-orthonormal, i.e., $\mathbf{Z}_t^\top \mathbf{B}\mathbf{Z}_t = \mathbf{I}$ (or equivalently, the left-hand side of Eq. (14) is the QR-factorization of the right-hand side in inner product $\langle , \rangle_{\mathbf{B}}$), and

$$\boldsymbol{\xi}_t = \widehat{\boldsymbol{\xi}}_t \mathbf{C}_t^{-1} \tag{18}$$

with $\mathbf{C}_t = \prod_{j=t}^{1} \mathbf{R}_j$ for $t \geq 1$ and $\mathbf{C}_0 = \mathbf{I}$.

Lemma 3.2 $\sin\theta(\mathbf{Z}_t, \mathbf{U}_q) = \Omega((\frac{\sqrt{\beta}}{\lambda_1^+})^t \sin\theta(\mathbf{Z}_0, \mathbf{U}_q))$, $\cos\theta(\mathbf{Z}_t, \mathbf{U}_q) = \Omega((\frac{\lambda_q^+}{\lambda_1^+})^t \cos\theta(\mathbf{Z}_0, \mathbf{U}_q))$.

Proof From the proof of Lemma 3.6 and with notations there, we have that

$$\begin{aligned}
\mathbf{U}_{-q}^\top \mathbf{B}\widehat{\mathbf{Z}}_t &= p_t(\boldsymbol{\Lambda}_{-q})\mathbf{U}_{-q}^\top \mathbf{B}\mathbf{Z}_0 + \boldsymbol{\Gamma}, \\
(\mathbf{U}_q^\top \mathbf{B}\widehat{\mathbf{Z}}_t)^\dagger &= (\mathbf{I} + (\mathbf{U}_q^\top \mathbf{B}\mathbf{Z}_0)^\dagger \boldsymbol{\Omega})^\top \mathbf{Q}\mathbf{P}^\top (\mathbf{I} + 2\mathrm{sym}(\mathbf{P}\mathbf{Q}^\top \boldsymbol{\Omega}^\top \mathbf{P}\boldsymbol{\Sigma}^{-1}\mathbf{P}^\top) \\
&\quad + \mathbf{P}\boldsymbol{\Sigma}^{-1}\mathbf{P}^\top \boldsymbol{\Omega}\boldsymbol{\Omega}^\top \mathbf{P}\boldsymbol{\Sigma}^{-1}\mathbf{P}^\top)^{-1} \mathbf{P}\boldsymbol{\Sigma}^{-1}\mathbf{P}^\top p_t^{-1}(\boldsymbol{\Lambda}_q),
\end{aligned}$$

where

$$\begin{aligned}
\boldsymbol{\Gamma} &= \textstyle\sum_{i=0}^{t-1} q_i(\boldsymbol{\Lambda}_{-q})\mathbf{U}_{-q}^\top \mathbf{B}\widehat{\boldsymbol{\xi}}_{t-i-1}, \\
\boldsymbol{\Omega} &= p_t^{-1}(\boldsymbol{\Lambda}_q) \textstyle\sum_{i=0}^{t-1} q_i(\boldsymbol{\Lambda}_q)\mathbf{U}_q^\top \mathbf{B}\widehat{\boldsymbol{\xi}}_{t-i-1}.
\end{aligned}$$

Then

$$\begin{aligned}
\|\mathbf{C}_t\|_2 \sin\theta(\mathbf{Z}_t, \mathbf{U}_q) &= \|\mathbf{U}_{-q}^\top \mathbf{B}\mathbf{Z}_t\|_2 \|\mathbf{C}_t\|_2 \\
&\geq \|\mathbf{U}_{-q}^\top \mathbf{B}\mathbf{Z}_t\mathbf{C}_t\|_2 = \|\mathbf{U}_{-q}^\top \mathbf{B}\widehat{\mathbf{Z}}_t\|_2 \\
&= \|p_t(\boldsymbol{\Lambda}_{-q})\mathbf{U}_{-q}^\top \mathbf{B}\mathbf{Z}_0 + \boldsymbol{\Gamma}\|_2 \\
&\geq \|p_t(\boldsymbol{\Lambda}_{-q})\mathbf{U}_{-q}^\top \mathbf{B}\mathbf{Z}_0\|_2 - \|\boldsymbol{\Gamma}\|_2 \\
&\geq \sigma_{\min}(p_t(\boldsymbol{\Lambda}_{-q}))\|\mathbf{U}_{-q}^\top \mathbf{B}\mathbf{Z}_0\|_2 - \sum_{i=0}^{t-1} \|q_i(\boldsymbol{\Lambda}_{-q})\|_2 \|\widehat{\boldsymbol{\xi}}_{t-i-1}\|_{\mathbf{B},2}
\end{aligned}$$

$$\begin{aligned}
&\geq (\sqrt{\beta})^t \min_{j>q} |\cos(t \arg\cos \frac{\lambda_j}{2\sqrt{\beta}})| \cdot \sin\theta(\mathbf{Z}_0, \mathbf{U}_q) \\
&\qquad - c\sum_{i=0}^{t-1} (i+1)(\sqrt{\beta})^i \frac{(\sqrt{\beta})^{t-i}\sin\theta(\mathbf{Z}_0, \mathbf{U}_q)}{(T-t+i+2)T} \\
&= c'(\sqrt{\beta})^t \sin\theta(\mathbf{Z}_0, \mathbf{U}_q),
\end{aligned}$$

where c, c' are sufficiently small positive constants, and in the last inequality we have used Lemma 3.7, Assumption (21) in the proof of Theorem 4.1, and that

$$p_t(\lambda_j) = \frac{(\lambda_j^+)^t + (\lambda_j^-)^t}{2} = (\sqrt{\beta})^t \cos(t \arg\cos \frac{\lambda_j}{2\sqrt{\beta}}),$$

for $\lambda_j \leq 2\sqrt{\beta}$. On the other hand, by Lemma 3.7 and Assumption (21), it holds that

$$\begin{aligned}
\|\boldsymbol{\Omega}\| &\leq \sum_{i=0}^{t-1} \|p_t^{-1}(\boldsymbol{\Lambda}_q) q_i(\boldsymbol{\Lambda}_q)\| \|\mathbf{U}_q^\top \mathbf{B}\widehat{\boldsymbol{\xi}}_{t-i-1}\|_2 \\
&\leq 2\sum_{i=0}^{t-1} (i+1)(\lambda_q^+)^{i-t} \frac{(\lambda_q^+)^{t-i}\cos\theta(\mathbf{Z}_0, \mathbf{U}_q)}{16(T-t+i+2)T} \\
&\leq \frac{1}{8}\cos\theta(\mathbf{Z}_0, \mathbf{U}_q).
\end{aligned}$$

By Lemma 3.7, Eq. (24) in the proof of Theorem 4.1, and the above inequality, we can write that

$$\begin{aligned}
\cos^{-1}\theta(\mathbf{Z}_t,\mathbf{U}_q) &= \|(\mathbf{U}_q^\top\mathbf{B}\mathbf{Z}_t)^\dagger\|_2 \\
&= \|\mathbf{C}_t(\mathbf{U}_q^\top\mathbf{B}\widehat{\mathbf{Z}}_t)^\dagger\|_2 \\
&\le \|\mathbf{C}_t\|_2\|(\mathbf{U}_q^\top\mathbf{B}\widehat{\mathbf{Z}}_t)^\dagger\|_2 \\
&\le \|\mathbf{C}_t\|_2\|(\mathbf{I}+(\mathbf{U}_q^\top\mathbf{B}\mathbf{Z}_0)^\dagger\boldsymbol{\Omega})^\top\mathbf{Q}\mathbf{P}^\top(\mathbf{I}+2\mathrm{sym}(\mathbf{P}\mathbf{Q}^\top\boldsymbol{\Omega}^\top \\
&\qquad \mathbf{P}\boldsymbol{\Sigma}^{-1}\mathbf{P}^\top)+\mathbf{P}\boldsymbol{\Sigma}^{-1}\mathbf{P}^\top\boldsymbol{\Omega}\boldsymbol{\Omega}^\top\mathbf{P}\boldsymbol{\Sigma}^{-1}\mathbf{P}^\top)^{-1}\mathbf{P}\boldsymbol{\Sigma}^{-1}\mathbf{P}^\top p_T^{-1}(\boldsymbol{\Lambda}_q)\|_2 \\
&\le \|\mathbf{C}_t\|_2\|p_T^{-1}(\boldsymbol{\Lambda}_q)\|_2\|\boldsymbol{\Sigma}^{-1}\|_2(1-2\|\boldsymbol{\Sigma}^{-1}\|_2\|\boldsymbol{\Omega}\|_2-(\|\boldsymbol{\Sigma}^{-1}\|_2\|\boldsymbol{\Omega}\|)^2)^{-1} \\
&\quad (1+\|(\mathbf{U}_q^\top\mathbf{B}\mathbf{Z}_0)^\dagger\|_2\|\boldsymbol{\Omega}\|_2) \\
&\le 8\|\mathbf{C}_t\|_2(\lambda_q^+)^{-t}\cos^{-1}\theta(\mathbf{Z}_0,\mathbf{U}_q).
\end{aligned}$$

Finally, by Lemma 3.8,

$$\begin{aligned}
\sin\theta(\mathbf{Z}_t,\mathbf{U}_q) &= \Omega((\frac{\sqrt{\beta}}{\lambda_1^+})^t\sin\theta(\mathbf{Z}_0,\mathbf{U}_q)), \\
\cos\theta(\mathbf{Z}_t,\mathbf{U}_q) &= \Omega((\frac{\lambda_q^+}{\lambda_1^+})^t\cos\theta(\mathbf{Z}_0,\mathbf{U}_q)).
\end{aligned}$$

□

Lemma 3.3 *$\widehat{\mathbf{Z}}_t=\mathbf{Z}_t\mathbf{C}_t$ holds for $t\ge 0$.*

Proof By Eq. (15), (17), and (18), we have

$$\mathbf{Z}_1=(\frac{1}{2}\mathbf{B}^{-1}\mathbf{A}\mathbf{Z}_0+\boldsymbol{\xi}_0)\mathbf{R}_1^{-1}=(\frac{1}{2}\mathbf{B}^{-1}\mathbf{A}\widehat{\mathbf{Z}}_0+\widehat{\boldsymbol{\xi}}_0)\mathbf{R}_1^{-1}=\widehat{\mathbf{Z}}_1\mathbf{R}_1^{-1}.$$

Thus, the target equation holds for $t=0,1$. Assume it holds for $(t-1),t$ and consider $\mathbf{X}_{t+1}$. By Eq. (14), (16), (18), and the hypothesis, we can write that

$$\begin{aligned}
\mathbf{Z}_{t+1} &= (\mathbf{B}^{-1}\mathbf{A}\mathbf{Z}_t-\beta\mathbf{Z}_{t-1}\mathbf{R}_t^{-1}+\boldsymbol{\xi}_t)\mathbf{R}_{t+1}^{-1} \\
&= (\mathbf{B}^{-1}\mathbf{A}\widehat{\mathbf{Z}}_t\mathbf{C}_t^{-1}-\beta\widehat{\mathbf{Z}}_{t-1}\mathbf{C}_{t-1}^{-1}\mathbf{R}_t^{-1}+\widehat{\boldsymbol{\xi}}_t\mathbf{C}_t^{-1})\mathbf{R}_{t+1}^{-1} \\
&= (\mathbf{B}^{-1}\mathbf{A}\widehat{\mathbf{Z}}_t\mathbf{C}_t^{-1}-\beta\widehat{\mathbf{Z}}_{t-1}\mathbf{C}_t^{-1}+\widehat{\boldsymbol{\xi}}_t\mathbf{C}_t^{-1})\mathbf{R}_{t+1}^{-1} \\
&= (\mathbf{B}^{-1}\mathbf{A}\widehat{\mathbf{Z}}_t-\beta\widehat{\mathbf{Z}}_{t-1}+\widehat{\boldsymbol{\xi}}_t)\mathbf{C}_t^{-1}\mathbf{R}_{t+1}^{-1} \\
&= \widehat{\mathbf{Z}}_{t+1}\mathbf{C}_{t+1}^{-1}.
\end{aligned}$$

By induction, it holds for all $t\ge 0$. □

Lemma 3.4 *$p_t(x)=\mathbf{a}_t\begin{bmatrix}\frac{1}{2}x\\1\end{bmatrix}=\frac{(x^+)^t+(x^-)^t}{2}$, $q_t(x)=\mathbf{a}_t\begin{bmatrix}x\\1\end{bmatrix}=\sum_{j=0}^t(x^+)^{t-j}(x^-)^j$, where $\mathbf{a}_t=\begin{bmatrix}0 & 1\end{bmatrix}\begin{bmatrix}x & -\beta\\1 & 0\end{bmatrix}^t$, and $x^\pm=\frac{x\pm\sqrt{x^2-4\beta}}{2}$ which is a conjugate pair when $|x|<2\sqrt{\beta}$.*

Proof By the definition, we can write that

$$\begin{bmatrix}p_{t+1}(x)\\p_t(x)\end{bmatrix}=\begin{bmatrix}xp_t(x)-\beta p_{t-1}(x)\\p_t(x)\end{bmatrix}=\begin{bmatrix}x & -\beta\\1 & 0\end{bmatrix}\begin{bmatrix}p_t(x)\\p_{t-1}(x)\end{bmatrix}=\cdots=\begin{bmatrix}x & -\beta\\1 & 0\end{bmatrix}^t\begin{bmatrix}p_1(x)\\p_0(x)\end{bmatrix}.$$

Thus, we have that

$$p_t(x)=\begin{bmatrix}0 & 1\end{bmatrix}\begin{bmatrix}x & -\beta\\1 & 0\end{bmatrix}^t\begin{bmatrix}\frac{1}{2}x\\1\end{bmatrix},\quad q_t(x)=\begin{bmatrix}0 & 1\end{bmatrix}\begin{bmatrix}x & -\beta\\1 & 0\end{bmatrix}^t\begin{bmatrix}x\\1\end{bmatrix}.$$

Note that the 2×2 matrix has the following Jordan decomposition[6]:

$$\begin{bmatrix} x & -\beta \\ 1 & 0 \end{bmatrix} = \begin{cases} \begin{bmatrix} x^+ & x^- \\ 1 & 1 \end{bmatrix} \begin{bmatrix} x^+ & 0 \\ 0 & x^- \end{bmatrix} \begin{bmatrix} x^+ & x^- \\ 1 & 1 \end{bmatrix}^{-1}, & |x| \neq 2\sqrt{\beta} \\ \begin{bmatrix} \frac{x}{2} & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} \frac{x}{2} & 1 \\ 0 & \frac{x}{2} \end{bmatrix} \begin{bmatrix} \frac{x}{2} & 1 \\ 1 & 0 \end{bmatrix}^{-1}, & |x| = 2\sqrt{\beta} \end{cases},$$

where $x^{\pm} = \frac{x \pm \sqrt{x^2 - 4\beta}}{2}$ are two eigenvalues of the 2×2 matrix which are equal if $|x| = 2\sqrt{\beta}$ and a conjugate pair if $|x| < 2\sqrt{\beta}$, and $|x^{\pm}| = \begin{cases} \text{sign}(x^{\pm})x^{\pm}, & |x| \geq 2\sqrt{\beta} \\ \sqrt{\beta}, & |x| \leq 2\sqrt{\beta} \end{cases}$. Note that

$$\begin{aligned} \begin{bmatrix} \frac{x}{2} & 1 \\ 0 & \frac{x}{2} \end{bmatrix}^t &= (\frac{x}{2}\mathbf{I} + \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix})^t = \sum_{j=0}^{t} \binom{t}{j} (\frac{x}{2})^{t-j} \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}^j \\ &= \binom{t}{0} (\frac{x}{2})^t \mathbf{I} + \binom{t}{1} (\frac{x}{2})^{t-1} \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} = (\frac{x}{2})^{t-1} \begin{bmatrix} \frac{x}{2} & t \\ 0 & \frac{x}{2} \end{bmatrix}. \end{aligned}$$

We thus have that

$$\begin{aligned} \begin{bmatrix} 0 & 1 \end{bmatrix} \begin{bmatrix} x & -\beta \\ 1 & 0 \end{bmatrix}^t &= \begin{cases} \begin{bmatrix} 0 & 1 \end{bmatrix} \begin{bmatrix} x^+ & x^- \\ 1 & 1 \end{bmatrix} \begin{bmatrix} x^+ & 0 \\ 0 & x^- \end{bmatrix}^t \begin{bmatrix} x^+ & x^- \\ 1 & 1 \end{bmatrix}^{-1}, & |x| \neq 2\sqrt{\beta} \\ \begin{bmatrix} 0 & 1 \end{bmatrix} \begin{bmatrix} \frac{x}{2} & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} \frac{x}{2} & 1 \\ 0 & \frac{x}{2} \end{bmatrix}^t \begin{bmatrix} \frac{x}{2} & 1 \\ 1 & 0 \end{bmatrix}^{-1}, & |x| = 2\sqrt{\beta} \end{cases} \\ &= \begin{cases} \frac{1}{x^+ - x^-} \begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} (x^+)^t & 0 \\ 0 & (x^-)^t \end{bmatrix} \begin{bmatrix} 1 & -x^- \\ -1 & x^+ \end{bmatrix}, & |x| \neq 2\sqrt{\beta} \\ \begin{bmatrix} 1 & 0 \end{bmatrix} (\frac{x}{2})^{t-1} \begin{bmatrix} \frac{x}{2} & t \\ 0 & \frac{x}{2} \end{bmatrix} (-1) \begin{bmatrix} 0 & -1 \\ -1 & \frac{x}{2} \end{bmatrix}, & |x| = 2\sqrt{\beta} \end{cases} \\ &= \begin{cases} \frac{1}{x^+ - x^-} \begin{bmatrix} (x^+)^t - (x^-)^t, & -(x^+)^t x^- + x^+ (x^-)^t \end{bmatrix}, & |x| \neq 2\sqrt{\beta} \\ (\frac{x}{2})^{t-1} \begin{bmatrix} t, & \frac{x}{2} - \frac{x}{2}t \end{bmatrix}, & |x| = 2\sqrt{\beta} \end{cases} \end{aligned}$$

In turn, we get that

$$\begin{aligned} p_t(x) &= \begin{cases} \frac{1}{x^+ - x^-} \begin{bmatrix} (x^+)^t - (x^-)^t, & -(x^+)^t x^- + x^+ (x^-)^t \end{bmatrix} \begin{bmatrix} \frac{x}{2} \\ 1 \end{bmatrix}, & |x| \neq 2\sqrt{\beta} \\ (\frac{x}{2})^{t-1} \begin{bmatrix} t, & \frac{x}{2} - \frac{x}{2}t \end{bmatrix} \begin{bmatrix} \frac{x}{2} \\ 1 \end{bmatrix}, & |x| = 2\sqrt{\beta} \end{cases} \\ &= \begin{cases} \frac{1}{x^+ - x^-}((\frac{x}{2} - x^-)(x^+)^t - (\frac{x}{2} - x^+)(x^-)^t), & |x| \neq 2\sqrt{\beta} \\ (\frac{x}{2})^{t-1}(\frac{x}{2}t + \frac{x}{2} - \frac{x}{2}t), & |x| = 2\sqrt{\beta} \end{cases} \\ &= \begin{cases} \frac{1}{2}((x^+)^t + (x^-)^t), & |x| \neq 2\sqrt{\beta} \\ (\frac{x}{2})^t, & |x| = 2\sqrt{\beta} \end{cases} = \frac{1}{2}((x^+)^t + (x^-)^t), \end{aligned}$$

[6] $x^{\pm}$ are the roots of $\det(\begin{bmatrix} x & -\beta \\ 1 & 0 \end{bmatrix} - \lambda\mathbf{I}) = 0$, i.e., $\lambda^2 - x\lambda + \beta = 0$. It is easy to check that $\begin{bmatrix} x & -\beta \\ 1 & 0 \end{bmatrix} \begin{bmatrix} x^{\pm} \\ 1 \end{bmatrix} = x^{\pm} \begin{bmatrix} x^{\pm} \\ 1 \end{bmatrix}$ holds. Thus, we have the eigenvalue decomposition when $x^+ \neq x^-$, where the eigenvector matrix is only non-singular but not orthogonal. When $x^+ = x^-$ which is $\frac{x}{2}$, two eigenvectors collapse into a single one and we need a generalized eigenvector which can be generated by solving $(\begin{bmatrix} x & -\beta \\ 1 & 0 \end{bmatrix} - x^+\mathbf{I})\mathbf{y} = \begin{bmatrix} x^+ \\ 1 \end{bmatrix}$. The equation is satisfied by $\mathbf{y} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$. We thus have $\begin{bmatrix} x & -\beta \\ 1 & 0 \end{bmatrix} \mathbf{y} = \begin{bmatrix} x^+ \\ 1 \end{bmatrix} + x^+\mathbf{y}$. Thus, we have the Jordan decomposition when $x^+ = x^-$.

$$
\begin{aligned}
q_t(x) &= \begin{cases} \frac{1}{x^+-x^-}\left[(x^+)^t-(x^-)^t, \quad -(x^+)^t x^- + x^+(x^-)^t\right]\begin{bmatrix} x \\ 1 \end{bmatrix}, & |x| \neq 2\sqrt{\beta} \\ (\frac{x}{2})^{t-1}\left[t, \quad \frac{x}{2}-\frac{x}{2}t\right]\begin{bmatrix} x \\ 1 \end{bmatrix}, & |x| = 2\sqrt{\beta} \end{cases} \\
&= \begin{cases} \frac{1}{x^+-x^-}((x-x^-)(x^+)^t-(x-x^+)(x^-)^t), & |x| \neq 2\sqrt{\beta} \\ (\frac{x}{2})^{t-1}(xt+\frac{x}{2}-\frac{x}{2}t), & |x| = 2\sqrt{\beta} \end{cases} \\
&= \begin{cases} \frac{1}{x^+-x^-}((x^+)^{t+1}-(x^-)^{t+1}), & |x| \neq 2\sqrt{\beta} \\ (\frac{x}{2})^{t-1}(\frac{x}{2}+\frac{x}{2}t), & |x| = 2\sqrt{\beta} \end{cases} \\
&= \begin{cases} \frac{1}{x^+-x^-}((x^+)^{t+1}-(x^-)^{t+1}), & |x| \neq 2\sqrt{\beta} \\ (t+1)(\frac{x}{2})^t, & |x| = 2\sqrt{\beta} \end{cases} = \sum_{j=0}^{t}(x^+)^{t-j}(x^-)^j.
\end{aligned}
$$

□

Lemma 3.5 $\widehat{\mathbf{Z}}_t = \mathbf{B}^{-\frac{1}{2}} p_t(\mathbf{B}^{-\frac{1}{2}}\mathbf{A}\mathbf{B}^{-\frac{1}{2}})\mathbf{B}^{\frac{1}{2}}\widehat{\mathbf{Z}}_0 + \sum_{j=0}^{t-1}\mathbf{B}^{-\frac{1}{2}} q_j(\mathbf{B}^{-\frac{1}{2}}\mathbf{A}\mathbf{B}^{-\frac{1}{2}})\mathbf{B}^{\frac{1}{2}}\widehat{\boldsymbol{\xi}}_{t-j-1}$.

Proof First, we have that

$$
\begin{aligned}
p_{t+1}(\mathbf{A}) &= \mathbf{A}p_t(\mathbf{A}) - \beta p_{t-1}(\mathbf{A}), \quad p_0(\mathbf{A}) = \mathbf{I}, \quad p_1(\mathbf{A}) = \frac{\mathbf{A}}{2}, \\
q_{t+1}(\mathbf{A}) &= \mathbf{A}q_t(\mathbf{A}) - \beta q_{t-1}(\mathbf{A}), \quad q_0(\mathbf{A}) = \mathbf{I}, \quad q_1(\mathbf{A}) = \mathbf{A}.
\end{aligned}
$$

For our purpose, we need to replace $\mathbf{A}$ with $\mathbf{B}^{-\frac{1}{2}}\mathbf{A}\mathbf{B}^{-\frac{1}{2}}$ in above equations. By Eq. (17) and initials of $p_t(\mathbf{B}^{-\frac{1}{2}}\mathbf{A}\mathbf{B}^{-\frac{1}{2}})$ and $q_t(\mathbf{B}^{-\frac{1}{2}}\mathbf{A}\mathbf{B}^{-\frac{1}{2}})$, the target equation holds for $t = 0, 1$. Assume it holds for $(t-1), t$ and consider $\widehat{\mathbf{Z}}_{t+1}$. By Eq. (16), the hypothesis, and two matrix polynomials' three-term recurrences as well as their initials above, we can write that

$$
\begin{aligned}
\widehat{\mathbf{Z}}_{t+1} &= \mathbf{B}^{-1}\mathbf{A}\widehat{\mathbf{Z}}_t - \beta\widehat{\mathbf{Z}}_{t-1} + \widehat{\boldsymbol{\xi}}_t \\
&= \mathbf{B}^{-1}\mathbf{A}\left(\mathbf{B}^{-\frac{1}{2}} p_t(\mathbf{B}^{-\frac{1}{2}}\mathbf{A}\mathbf{B}^{-\frac{1}{2}})\mathbf{B}^{\frac{1}{2}}\widehat{\mathbf{Z}}_0 + \sum_{j=0}^{t-1}\mathbf{B}^{-\frac{1}{2}} q_j(\mathbf{B}^{-\frac{1}{2}}\mathbf{A}\mathbf{B}^{-\frac{1}{2}})\mathbf{B}^{\frac{1}{2}}\widehat{\boldsymbol{\xi}}_{t-j-1}\right) \\
&\quad -\beta\left(\mathbf{B}^{-\frac{1}{2}} p_{t-1}(\mathbf{B}^{-\frac{1}{2}}\mathbf{A}\mathbf{B}^{-\frac{1}{2}})\mathbf{B}^{\frac{1}{2}}\widehat{\mathbf{Z}}_0 + \sum_{j=0}^{t-2}\mathbf{B}^{-\frac{1}{2}} q_j(\mathbf{B}^{-\frac{1}{2}}\mathbf{A}\mathbf{B}^{-\frac{1}{2}})\mathbf{B}^{\frac{1}{2}}\widehat{\boldsymbol{\xi}}_{t-j-2}\right) + \widehat{\boldsymbol{\xi}}_t \\
&= \mathbf{B}^{-1}\mathbf{A}\Bigg(\mathbf{B}^{-\frac{1}{2}} p_t(\mathbf{B}^{-\frac{1}{2}}\mathbf{A}\mathbf{B}^{-\frac{1}{2}})\mathbf{B}^{\frac{1}{2}}\widehat{\mathbf{Z}}_0 + \mathbf{B}^{-\frac{1}{2}} q_0(\mathbf{B}^{-\frac{1}{2}}\mathbf{A}\mathbf{B}^{-\frac{1}{2}})\mathbf{B}^{\frac{1}{2}}\widehat{\boldsymbol{\xi}}_{t-1} \\
&\qquad\qquad + \sum_{j=1}^{t-1}\mathbf{B}^{-\frac{1}{2}} q_j(\mathbf{B}^{-\frac{1}{2}}\mathbf{A}\mathbf{B}^{-\frac{1}{2}})\mathbf{B}^{\frac{1}{2}}\widehat{\boldsymbol{\xi}}_{t-j-1}\Bigg) \\
&\quad -\beta\left(\mathbf{B}^{-\frac{1}{2}} p_{t-1}(\mathbf{B}^{-\frac{1}{2}}\mathbf{A}\mathbf{B}^{-\frac{1}{2}})\mathbf{B}^{\frac{1}{2}}\widehat{\mathbf{Z}}_0 + \sum_{j=1}^{t-1}\mathbf{B}^{-\frac{1}{2}} q_{j-1}(\mathbf{B}^{-\frac{1}{2}}\mathbf{A}\mathbf{B}^{-\frac{1}{2}})\mathbf{B}^{\frac{1}{2}}\widehat{\boldsymbol{\xi}}_{t-j-1}\right) + \widehat{\boldsymbol{\xi}}_t \\
&= \left(\mathbf{B}^{-1}\mathbf{A}\mathbf{B}^{-\frac{1}{2}} p_t(\mathbf{B}^{-\frac{1}{2}}\mathbf{A}\mathbf{B}^{-\frac{1}{2}}) - \beta\mathbf{B}^{-\frac{1}{2}} p_{t-1}(\mathbf{B}^{-\frac{1}{2}}\mathbf{A}\mathbf{B}^{-\frac{1}{2}})\right)\mathbf{B}^{\frac{1}{2}}\widehat{\mathbf{Z}}_0 + \mathbf{B}^{-1}\mathbf{A}\widehat{\boldsymbol{\xi}}_{t-1} \\
&\quad + \sum_{j=1}^{t-1}\left(\mathbf{B}^{-1}\mathbf{A}\mathbf{B}^{-\frac{1}{2}} q_j(\mathbf{B}^{-\frac{1}{2}}\mathbf{A}\mathbf{B}^{-\frac{1}{2}}) - \beta\mathbf{B}^{-\frac{1}{2}} q_{j-1}(\mathbf{B}^{-\frac{1}{2}}\mathbf{A}\mathbf{B}^{-\frac{1}{2}})\right)\mathbf{B}^{\frac{1}{2}}\widehat{\boldsymbol{\xi}}_{t-j-1} + \widehat{\boldsymbol{\xi}}_t \\
&= \mathbf{B}^{-\frac{1}{2}} p_{t+1}(\mathbf{B}^{-\frac{1}{2}}\mathbf{A}\mathbf{B}^{-\frac{1}{2}})\mathbf{B}^{\frac{1}{2}}\widehat{\mathbf{Z}}_0 + \mathbf{B}^{-\frac{1}{2}} q_1(\mathbf{A})\mathbf{B}^{\frac{1}{2}}\widehat{\boldsymbol{\xi}}_{t-1} \\
&\quad + \sum_{j=1}^{t-1}\mathbf{B}^{-\frac{1}{2}} q_{j+1}(\mathbf{B}^{-\frac{1}{2}}\mathbf{A}\mathbf{B}^{-\frac{1}{2}})\mathbf{B}^{\frac{1}{2}}\widehat{\boldsymbol{\xi}}_{t-j-1} + \mathbf{B}^{-\frac{1}{2}} q_0(\mathbf{A})\mathbf{B}^{\frac{1}{2}}\widehat{\boldsymbol{\xi}}_t
\end{aligned}
$$

$$= \mathbf{B}^{-\frac{1}{2}} p_{t+1}(\mathbf{B}^{-\frac{1}{2}}\mathbf{A}\mathbf{B}^{-\frac{1}{2}})\widehat{\mathbf{Z}}_0 + \sum_{j=0}^{t} \mathbf{B}^{-\frac{1}{2}} q_j(\mathbf{B}^{-\frac{1}{2}}\mathbf{A}\mathbf{B}^{-\frac{1}{2}})\mathbf{B}^{\frac{1}{2}}\widehat{\boldsymbol{\xi}}_{t-j},$$

which completes the proof, by induction. □

Lemma 3.6 *Let $\mathbf{U}_q^\top \mathbf{B}\mathbf{Z}_0 = \mathbf{P}\boldsymbol{\Sigma}\mathbf{Q}^\top$ be the SVD of $\mathbf{U}_q^\top \mathbf{Z}_0$, where $\mathbf{P} \in \mathbb{R}^{q\times q}$ is orthogonal, $\boldsymbol{\Lambda} \in \mathbb{R}^{q\times q}$ is diagonal, and $\mathbf{Q} \in \mathbb{R}^{p\times q}$ is column-orthonormal. It holds that*

$$\begin{aligned}
\mathbf{U}_{-q}^\top \mathbf{B}\widehat{\mathbf{Z}}_T &= p_T(\boldsymbol{\Lambda}_{-q})\mathbf{U}_{-q}^\top \mathbf{B}\mathbf{Z}_0 + \boldsymbol{\Gamma}, \\
(\mathbf{U}_q^\top \mathbf{B}\widehat{\mathbf{X}}_T)^\dagger \begin{bmatrix}\mathbf{I}_k & \mathbf{0}\end{bmatrix}^\top &= (\mathbf{I} + (\mathbf{U}_q^\top \mathbf{B}\mathbf{X}_0)^\dagger \boldsymbol{\Omega})^\top \mathbf{Q}\mathbf{P}^\top (\mathbf{I} + 2\mathrm{sym}(\mathbf{P}\mathbf{Q}^\top \boldsymbol{\Omega}^\top \mathbf{P}\boldsymbol{\Sigma}^{-1}\mathbf{P}^\top) \\
&\quad + \mathbf{P}\boldsymbol{\Sigma}^{-1}\mathbf{P}^\top \boldsymbol{\Omega}\boldsymbol{\Omega}^\top \mathbf{P}\boldsymbol{\Sigma}^{-1}\mathbf{P}^\top)^{-1}\mathbf{P}\boldsymbol{\Sigma}^{-1}\mathbf{P}^\top \begin{bmatrix}p_T^{-1}(\boldsymbol{\Lambda}_k) & \mathbf{0}\end{bmatrix}^\top,
\end{aligned}$$

where $\mathrm{sym}(\cdot)$ *extracts the symmetric part of a matrix and*

$$\boldsymbol{\Gamma} = \sum_{t=0}^{T-1} q_t(\boldsymbol{\Lambda}_{-q})\mathbf{U}_{-q}^\top \mathbf{B}\widehat{\boldsymbol{\xi}}_{T-t-1}, \quad \boldsymbol{\Omega} = p_T^{-1}(\boldsymbol{\Lambda}_q) \sum_{t=0}^{T-1} q_t(\boldsymbol{\Lambda}_q)\mathbf{U}_q^\top \mathbf{B}\widehat{\boldsymbol{\xi}}_{T-t-1}.$$

Proof First, it holds for any $1 \le j \le n$ that

$$p_t(\mathbf{B}^{-\frac{1}{2}}\mathbf{A}\mathbf{B}^{-\frac{1}{2}}) = \mathbf{B}^{\frac{1}{2}} \begin{bmatrix}\mathbf{U}_j & \mathbf{U}_{-j}\end{bmatrix} \mathrm{diag}(p_t(\boldsymbol{\Lambda}_j), p_t(\boldsymbol{\Lambda}_{-j})) \begin{bmatrix}\mathbf{U}_j & \mathbf{U}_{-j}\end{bmatrix}^\top \mathbf{B}^{\frac{1}{2}}.$$

We then have that

$$\begin{aligned}
\mathbf{U}_{-q}^\top \mathbf{B}\widehat{\mathbf{Z}}_T &= \mathbf{U}_{-q}^\top \mathbf{B}\Big(\mathbf{B}^{-\frac{1}{2}} p_T(\mathbf{B}^{-\frac{1}{2}}\mathbf{A}\mathbf{B}^{-\frac{1}{2}})\mathbf{B}^{\frac{1}{2}}\mathbf{Z}_0 + \sum_{t=0}^{T-1} \mathbf{B}^{-\frac{1}{2}} q_t(\mathbf{B}^{-\frac{1}{2}}\mathbf{A}\mathbf{B}^{-\frac{1}{2}})\mathbf{B}^{\frac{1}{2}}\widehat{\boldsymbol{\xi}}_{T-t-1}\Big) \\
&= p_T(\boldsymbol{\Lambda}_{-q})\mathbf{U}_{-q}^\top \mathbf{B}\mathbf{Z}_0 + \boldsymbol{\Gamma}.
\end{aligned}$$

Second, we can write that

$$\begin{aligned}
(\mathbf{U}_q^\top \mathbf{B}\widehat{\mathbf{Z}}_T)^\dagger &= \Big(\mathbf{U}_q^\top \mathbf{B}\big(\mathbf{B}^{-\frac{1}{2}} p_T(\mathbf{B}^{-\frac{1}{2}}\mathbf{A}\mathbf{B}^{-\frac{1}{2}})\mathbf{B}^{\frac{1}{2}}\mathbf{Z}_0 + \sum_{t=0}^{T-1} \mathbf{B}^{-\frac{1}{2}} q_t(\mathbf{B}^{-\frac{1}{2}}\mathbf{A}\mathbf{B}^{-\frac{1}{2}})\mathbf{B}^{\frac{1}{2}}\widehat{\boldsymbol{\xi}}_{T-t-1}\big)\Big)^\dagger \\
&= \Big(p_T(\boldsymbol{\Lambda}_q)\mathbf{U}_q^\top \mathbf{B}\mathbf{Z}_0 + \sum_{t=0}^{T-1} q_t(\boldsymbol{\Lambda}_q)\mathbf{U}_q^\top \mathbf{B}\widehat{\boldsymbol{\xi}}_{T-t-1}\Big)^\dagger \\
&= \Big(p_T(\boldsymbol{\Lambda}_q)\mathbf{U}_q^\top \mathbf{B}\mathbf{Z}_0 + \sum_{t=0}^{T-1} q_t(\boldsymbol{\Lambda}_q)\mathbf{U}_q^\top \mathbf{B}\widehat{\boldsymbol{\xi}}_{T-t-1}\Big)^\top \Big(\big(p_T(\boldsymbol{\Lambda}_q)\mathbf{U}_q^\top \mathbf{B}\mathbf{Z}_0 + \sum_{t=0}^{T-1} q_t(\boldsymbol{\Lambda}_q)\mathbf{U}_q^\top \mathbf{B}\widehat{\boldsymbol{\xi}}_{T-t-1}\big) \\
&\quad \big(p_T(\boldsymbol{\Lambda}_q)\mathbf{U}_q^\top \mathbf{B}\mathbf{Z}_0 + \sum_{t=0}^{T-1} q_t(\boldsymbol{\Lambda}_q)\mathbf{U}_q^\top \mathbf{B}\widehat{\boldsymbol{\xi}}_{T-t-1}\big)^\top\Big)^{-1} \\
&= (\mathbf{U}_q^\top \mathbf{B}\mathbf{Z}_0 + \boldsymbol{\Omega})^\top ((\mathbf{U}_q^\top \mathbf{B}\mathbf{Z}_0 + \boldsymbol{\Omega})(\mathbf{U}_q^\top \mathbf{B}\mathbf{Z}_0 + \boldsymbol{\Omega})^\top)^{-1} p_T^{-1}(\boldsymbol{\Lambda}_q) \\
&= (\mathbf{U}_q^\top \mathbf{B}\mathbf{Z}_0 + \boldsymbol{\Omega})^\top (\mathbf{U}_q^\top \mathbf{B}\mathbf{Z}_0 (\mathbf{U}_q^\top \mathbf{B}\mathbf{Z}_0)^\top + 2\mathrm{sym}(\mathbf{U}_q^\top \mathbf{B}\mathbf{Z}_0 \boldsymbol{\Omega}^\top) + \boldsymbol{\Omega}\boldsymbol{\Omega}^\top)^{-1} p_T^{-1}(\boldsymbol{\Lambda}_q)
\end{aligned}$$

where the first equality is by Lemma 3.5 and the third equality is by the definition of the matrix pseudo-inverse. Let $\boldsymbol{\Xi} = \mathbf{U}_q^\top \mathbf{B}\mathbf{Z}_0 (\mathbf{U}_q^\top \mathbf{B}\mathbf{Z}_0)^\top$ and note that $\mathbf{U}_q^\top \mathbf{B}\mathbf{Z}_0 + \boldsymbol{\Omega} = \mathbf{U}_q^\top \mathbf{B}\mathbf{Z}_0 + \mathbf{U}_q^\top \mathbf{B}\mathbf{Z}_0 (\mathbf{U}_q^\top \mathbf{B}\mathbf{Z}_0)^\dagger \boldsymbol{\Omega}$. We then have that

$$\begin{aligned}
(\mathbf{U}_q^\top \mathbf{B}\widehat{\mathbf{Z}}_T)^\dagger &= (\mathbf{I} + (\mathbf{U}_q^\top \mathbf{B}\mathbf{Z}_0)^\dagger \boldsymbol{\Omega})^\top (\mathbf{U}_q^\top \mathbf{B}\mathbf{Z}_0)^\top \boldsymbol{\Xi}^{-1/2} (\mathbf{I} + 2\boldsymbol{\Xi}^{-1/2}\mathrm{sym}(\mathbf{U}_q^\top \mathbf{B}\mathbf{Z}_0 \boldsymbol{\Omega}^\top)\boldsymbol{\Xi}^{-1/2} \\
&\quad + \boldsymbol{\Xi}^{-1/2}\boldsymbol{\Omega}\boldsymbol{\Omega}^\top \boldsymbol{\Xi}^{-1/2})^{-1} \boldsymbol{\Xi}^{-1/2} p_T^{-1}(\boldsymbol{\Lambda}_q) \\
&= (\mathbf{I} + (\mathbf{U}_q^\top \mathbf{B}\mathbf{Z}_0)^\dagger \boldsymbol{\Omega})^\top \mathbf{Q}\mathbf{P}^\top (\mathbf{I} + 2\mathrm{sym}(\mathbf{P}\mathbf{Q}^\top \boldsymbol{\Omega}^\top \mathbf{P}\boldsymbol{\Lambda}^{-1}\mathbf{P}^\top) \\
&\quad + \mathbf{P}\boldsymbol{\Sigma}^{-1}\mathbf{P}^\top \boldsymbol{\Omega}\boldsymbol{\Omega}^\top \mathbf{P}\boldsymbol{\Sigma}^{-1}\mathbf{P}^\top)^{-1}\mathbf{P}\boldsymbol{\Sigma}^{-1}\mathbf{P}^\top p_T^{-1}(\boldsymbol{\Lambda}_q).
\end{aligned}$$

The proof completes by noting that

$$\begin{bmatrix}\mathbf{I}_k & \mathbf{0}\end{bmatrix}^\top p_T^{-1}(\boldsymbol{\Lambda}_q) = \begin{bmatrix}p_T^{-1}(\boldsymbol{\Lambda}_k) & \mathbf{0}\end{bmatrix}^\top.$$

□

Lemma 3.7

$$\|p_T^{-1}(\mathbf{\Lambda}_k)\| \le 2(\lambda_k^+)^{-T}, \quad \|q_t(\mathbf{\Lambda}_{-q})\| \le (t+1)(\sqrt{\beta})^t,$$
$$\|p_T(\mathbf{\Lambda}_{-q})\| \le (\sqrt{\beta})^T, \quad \|p_T^{-1}(\mathbf{\Lambda}_q)q_t(\mathbf{\Lambda}_q)\| \le 2(t+1)(\lambda_q^+)^{t-T}.$$

Proof Since $\lambda_1 \ge \cdots \ge \lambda_k \ge \cdots \ge \lambda_q > 2\sqrt{\beta}$ by the assumption of Theorem 3.1, then by Lemma 3.4

$$\|p_T^{-1}(\mathbf{\Lambda}_k)\| = \max_{1\le m\le k} \frac{2}{(\lambda_m^+)^T + (\lambda_m^-)^T} \le \frac{2}{(\lambda_k^+)^T}.$$

Since $2\sqrt{\beta} \ge \lambda_{q+1} \ge \cdots \ge \lambda_n$ by the assumption of Theorem 3.1, then by Lemma 3.4

$$\begin{aligned}\|q_t(\mathbf{\Lambda}_{-q})\| &= \max_{q+1\le m\le n} |\textstyle\sum_{j=0}^t (\lambda_m^+)^{t-j}(\lambda_m^-)^j| \\ &\le \max_{q+1\le m\le n} \textstyle\sum_{j=0}^t |\lambda_m^+|^{t-j}|\lambda_m^-|^j \\ &= \max_{q+1\le m\le n} \textstyle\sum_{j=0}^t (\sqrt{\beta})^{t-j}(\sqrt{\beta})^j = (t+1)(\sqrt{\beta})^t.\end{aligned}$$

Similarly, we have that

$$\begin{aligned}\|p_T(\mathbf{\Lambda}_{-q})\| &= \max_{q+1\le m\le n} |(\lambda_m^+)^T + (\lambda_m^-)^T|/2 \\ &\le \max_{q+1\le m\le n} (|\lambda_m^+|^T + |\lambda_m^-|^T)/2 \\ &= \max_{q+1\le m\le n} ((\sqrt{\beta})^T + (\sqrt{\beta})^T)/2 = (\sqrt{\beta})^T,\end{aligned}$$

and

$$\begin{aligned}\|p_T^{-1}(\mathbf{\Lambda}_q)q_t(\mathbf{\Lambda}_q)\| &= \max_{1\le m\le q} \frac{\sum_{j=0}^t (\lambda_m^+)^{t-j}(\lambda_m^-)^j}{\frac{(\lambda_m^+)^T + (\lambda_m^-)^T}{2}} \\ &\le 2\max_{1\le m\le q} \textstyle\sum_{j=0}^t (\lambda_m^+)^t/(\lambda_m^+)^T \\ &= 2\max_{1\le m\le q} (t+1)(\lambda_m^+)^t/(\lambda_m^+)^T = 2(t+1)(\lambda_q^+)^{t-T}.\end{aligned}$$

□

Lemma 3.8 $\|\mathbf{C}_t\| = \Theta((\lambda_1^+)^t)$.

Proof We can bound $\|\mathbf{C}_t\|$ based on the connection defined in Eq. (18). By Lemmas 3.3 and 3.5, we have that

$$\begin{aligned}\mathbf{Z}_t\mathbf{C}_t &= \mathbf{B}^{-\frac{1}{2}}p_t(\mathbf{B}^{-\frac{1}{2}}\mathbf{A}\mathbf{B}^{-\frac{1}{2}})\mathbf{B}^{\frac{1}{2}}\mathbf{Z}_0 + \sum_{j=0}^{t-1}\mathbf{B}^{-\frac{1}{2}}q_j(\mathbf{B}^{-\frac{1}{2}}\mathbf{A}\mathbf{B}^{-\frac{1}{2}})\mathbf{B}^{\frac{1}{2}}\widehat{\boldsymbol{\xi}}_{t-j-1} \\ &= \mathbf{U}_n p_t(\mathbf{\Lambda}_n)\mathbf{U}_n^\top\mathbf{B}\mathbf{Z}_0 + \sum_{j=0}^{t-1}\mathbf{U}_n q_j(\mathbf{\Lambda}_n)\mathbf{U}_n^\top\mathbf{B}\widehat{\boldsymbol{\xi}}_{t-j-1}.\end{aligned}$$

Thus, by Lemma 3.4 and Eq. (8) in the main text which in the setting of generalized eigenspace computation is

$$\begin{cases}\|\widehat{\boldsymbol{\xi}}_t\|_{\mathbf{B},2} = O(\frac{1}{(T-t+1)T}(\sqrt{\beta})^{t+1}\sin\theta(\mathbf{Z}_0,\mathbf{U}_q)) \\ \|\mathbf{U}_q^\top\mathbf{B}\widehat{\boldsymbol{\xi}}_t\|_2 = O(\frac{1}{(T-t+1)T}(\lambda_q^+)^{t+1}\cos\theta(\mathbf{Z}_0,\mathbf{U}_q))\end{cases},$$

it holds that

$$\begin{aligned}\|\mathbf{C}_t\|_2 &= \|\mathbf{Z}_t\mathbf{C}_t\|_{\mathbf{B},2} = \|p_t(\mathbf{\Lambda}_n)\mathbf{U}_n^\top\mathbf{B}\mathbf{Z}_0 + \textstyle\sum_{j=0}^{t-1} q_j(\mathbf{\Lambda}_n)\mathbf{U}_n^\top\mathbf{B}\widehat{\boldsymbol{\xi}}_{t-j-1}\|_2 \\ &\le \|p_t(\mathbf{\Lambda}_n)\|_2 + \textstyle\sum_{j=0}^{t-1}\|q_j(\mathbf{\Lambda}_n)\|_2\|\widehat{\boldsymbol{\xi}}_{t-j-1}\|_{\mathbf{B},2}\end{aligned}$$

Algorithm 2 ANPM for Generalized Eigenspace Computation under $p \geq k$.

Input: real matrices $(\mathbf{A}, \mathbf{B}) \in \mathbb{R}^{n\times n} \times \mathbb{R}^{n\times n}$ with $\mathbf{A} \succcurlyeq \mathbf{0}, \mathbf{B} \succ \mathbf{0}$, momentum parameter $\beta > 0$, target rank k, iteration rank $p \geq k$, iteration number T, a subroutine $\mathrm{GS}_{\mathbf{B}}$ that performs modified Gram-Schmidt process with inner product $\langle , \rangle_{\mathbf{B}}$.
Output: approximate top-k eigenspace spanned by the first k columns of $\mathbf{X}_T$.
Sample an entry-wise i.i.d. standard Gaussian matrix $\mathbf{G} \in \mathbb{R}^{n\times p}$
$\mathbf{Z}_{-1} = \mathbf{0}$ and $\mathbf{Z}_0 = \mathrm{GS}_{\mathbf{B}}(\mathbf{G})$
for $t = 0, 1, \cdots, T-1$ **do**
  $\mathbf{H}_t = (\mathbf{Z}_t^\top \mathbf{B}\mathbf{Z}_t)^{-1}(\mathbf{Z}_t^\top \mathbf{A}\mathbf{Z}_t)$
  $\tilde{\mathbf{Z}}_{t+1} \approx \arg\min_{\mathbf{Z}} \mathrm{tr}(\frac{1}{2}\mathbf{Z}^\top\mathbf{B}\mathbf{Z} - \mathbf{Z}^\top\mathbf{A}\mathbf{Z}_t)$ with warm start $\mathbf{Z}_t\mathbf{H}_t$
  if $t = 0$ **then**
    $\mathbf{Z}_{t+1} = \mathrm{GS}_{\mathbf{B}}(\frac{1}{2}\tilde{\mathbf{Z}}_{t+1})$ such that $\mathbf{Z}_{t+1}\mathbf{R}_{t+1} = \frac{1}{2}\tilde{\mathbf{Z}}_{t+1}$
  else
    $\mathbf{Z}_{t+1} = \mathrm{GS}_{\mathbf{B}}(\tilde{\mathbf{Z}}_{t+1} - \beta\mathbf{Z}_{t-1}\mathbf{R}_t^{-1})$ such that $\mathbf{Z}_{t+1}\mathbf{R}_{t+1} = \tilde{\mathbf{Z}}_{t+1} - \beta\mathbf{Z}_{t-1}\mathbf{R}_t^{-1}$
  end if
end for

$$
\begin{aligned}
&= \max_{1\leq m\leq n} \frac{|(\lambda_m^+)^t + (\lambda_m^-)^t|}{2} + \textstyle\sum_{j=0}^{t-1} \|\widehat{\boldsymbol{\xi}}_{t-j-1}\|_{\mathbf{B},2} \cdot \max_{1\leq m\leq n} |\sum_{s=0}^{j} (\lambda_m^+)^{j-s}(\lambda_m^-)^s| \\
&\leq (\lambda_1^+)^t + \tfrac{1}{16}\textstyle\sum_{j=0}^{t-1} \frac{(\lambda_q^+)^{t-j}\cos\theta(\mathbf{Z}_0,\mathbf{U}_q)}{(T-t+j+2)T}(j+1)(\lambda_1^+)^j \\
&\leq (\lambda_1^+)^t + \tfrac{(\lambda_1^+)^t \cos\theta(\mathbf{Z}_0,\mathbf{U}_q)}{16T}\textstyle\sum_{j=0}^{t-1}\frac{j+1}{T-t+j+2} \leq 2(\lambda_1^+)^t.
\end{aligned}
$$

On the other hand, it holds that

$$
\begin{aligned}
\|\mathbf{C}_t\|_2 &= \|p_t(\boldsymbol{\Lambda}_n)\mathbf{U}_n^\top\mathbf{B}\mathbf{Z}_0 + \textstyle\sum_{j=0}^{t-1} q_j(\boldsymbol{\Lambda}_n)\mathbf{U}_n^\top\mathbf{B}\widehat{\boldsymbol{\xi}}_{t-j-1}\|_2 \\
&\geq \sigma_{\min}(\mathbf{U}_n^\top\mathbf{B}\mathbf{Z}_0)\|p_t(\boldsymbol{\Lambda}_n)\|_2 - \textstyle\sum_{j=0}^{t-1}\|q_j(\boldsymbol{\Lambda}_n)\|_2\|\widehat{\boldsymbol{\xi}}_{t-j-1}\|_{\mathbf{B},2} \\
&= \max_{1\leq m\leq n} \frac{|(\lambda_m^+)^t + (\lambda_m^-)^t|}{2} - \textstyle\sum_{j=0}^{t-1} \|\widehat{\boldsymbol{\xi}}_{t-j-1}\|_{\mathbf{B},2} \cdot \max_{1\leq m\leq n} |\sum_{s=0}^{j} (\lambda_m^+)^{j-s}(\lambda_m^-)^s| \\
&= \max\{\max_{m\leq q} \frac{(\lambda_m^+)^t + (\lambda_m^-)^t}{2}, \max_{m>q} (\sqrt{\beta})^t \cos(t\, arccos\frac{\lambda_m}{2\sqrt{\beta}})\} \\
&\quad - \textstyle\sum_{j=0}^{t-1} \|\widehat{\boldsymbol{\xi}}_{t-j-1}\|_{\mathbf{B},2} \cdot \max_{1\leq m\leq n} |\sum_{s=0}^{j} (\lambda_m^+)^{j-s}(\lambda_m^-)^s| \\
&= \frac{(\lambda_1^+)^t + (\lambda_1^-)^t}{2} - \textstyle\sum_{j=0}^{t-1} \|\widehat{\boldsymbol{\xi}}_{t-j-1}\|_{\mathbf{B},2} \cdot \max_{1\leq m\leq n} |\sum_{s=0}^{j} (\lambda_m^+)^{j-s}(\lambda_m^-)^s| \\
&\geq \frac{(\lambda_1^+)^t}{2} - \textstyle\sum_{j=0}^{t-1} \|\widehat{\boldsymbol{\xi}}_{t-j-1}\|_{\mathbf{B},2} \cdot \max_{1\leq m\leq n} |\sum_{s=0}^{j} (\lambda_m^+)^{j-s}(\lambda_m^-)^s| \\
&\geq \frac{1}{2}(\lambda_1^+)^t - \tfrac{1}{16}\textstyle\sum_{j=0}^{t-1} \frac{(\lambda_q^+)^{t-j}\cos\theta(\mathbf{Z}_0,\mathbf{U}_q)}{(T-t+j+2)T}(j+1)(\lambda_1^+)^j \\
&\geq \frac{1}{2}(\lambda_1^+)^t - \tfrac{(\lambda_1^+)^t\cos\theta(\mathbf{Z}_0,\mathbf{U}_q)}{16T}\textstyle\sum_{j=0}^{t-1}\frac{j+1}{T-t+j+2} \geq \frac{7}{16}(\lambda_1^+)^t.
\end{aligned}
$$

Thus, we can write $\|\mathbf{C}_t\|_2 = \Theta((\lambda_1^+)^t)$. □

Theorem 4.1 *Let $k \leq q \leq p$ and assume that $\lambda_q > 2\sqrt{\beta} \geq \lambda_{q+1}$ and $\lambda_n \geq 0$ for a pair of $n \times n$ real symmetric matrices $(\mathbf{A}, \mathbf{B})$ with $\mathbf{B} \succ \mathbf{0}$. After Algorithm 2 or the update below*

$$\mathbf{Z}_{t+1}\mathbf{R}_{t+1} = \mathbf{B}^{-1}\mathbf{A}\mathbf{Z}_t - \beta\mathbf{Z}_{t-1}\mathbf{R}_t^{-1} + \boldsymbol{\xi}_t$$

runs for $T = O(\frac{1}{\sqrt{\rho}}\log\frac{\tan\theta_0}{\epsilon})$ iterations, we have that $\sin\theta(\mathbf{Z}_T, \mathbf{U}_k) < \epsilon$ in time complexity

$$O(\mathrm{nnz}(\mathbf{B})p\sqrt{\frac{\kappa(\mathbf{B})}{\rho}}(\log\frac{1}{\cos\theta_0}\log\frac{\gamma}{\rho\cos\theta_0} + \log\frac{1}{\epsilon}\log\frac{\gamma}{\rho}) + \frac{\mathrm{nnz}(\mathbf{A})p + \mathrm{nnz}(\mathbf{B})p^2}{\sqrt{\rho}}\log\frac{1}{\epsilon\cos\theta_0}),$$

where $\rho = \frac{\lambda_k - 2\sqrt{\beta}}{\lambda_k}$, $\gamma = \frac{\lambda_1}{\lambda_k}$, $\kappa(\mathbf{B})$ *represents the condition number of* $\mathbf{B}$, $\mathrm{nnz}(\cdot)$ *represents the number of nonzero entries in a matrix, and* $\theta_0 = \theta(\mathbf{Z}_0, \mathbf{U}_q)$.

Proof Since the proof is quite similar to that of Theorem 3.1 in the main text, we focus on the differences. By Lemma 3.3 and 3.5, we can express $\mathbf{Z}_t$ in a closed form:

$$\mathbf{Z}_t = (\mathbf{B}^{-\frac{1}{2}} p_t(\mathbf{B}^{-\frac{1}{2}}\mathbf{A}\mathbf{B}^{-\frac{1}{2}})\mathbf{B}^{\frac{1}{2}}\mathbf{Z}_0 + \sum_{j=0}^{t-1} \mathbf{B}^{-\frac{1}{2}} q_j(\mathbf{B}^{-\frac{1}{2}}\mathbf{A}\mathbf{B}^{-\frac{1}{2}})\mathbf{B}^{\frac{1}{2}}\widehat{\boldsymbol{\xi}}_{t-j-1})\mathbf{C}_t^{-1}. \tag{19}$$

Now consider

$$h_T = \|(\mathbf{U}_{-q}^\top \mathbf{B}\mathbf{Z}_T)(\mathbf{U}_q^\top \mathbf{B}\mathbf{Z}_T)^\dagger \begin{bmatrix}\mathbf{I}_k & \mathbf{0}\end{bmatrix}^\top \| = \|(\mathbf{U}_{-q}^\top \mathbf{B}\widehat{\mathbf{Z}}_T)(\mathbf{U}_q^\top \mathbf{B}\widehat{\mathbf{Z}}_T)^\dagger \begin{bmatrix}\mathbf{I}_k & \mathbf{0}\end{bmatrix}^\top \|.$$

By Lemma 3.6, we can write that

$$\begin{aligned} h_T &= \|(p_T(\boldsymbol{\Lambda}_{-q})\mathbf{U}_{-q}^\top \mathbf{B}\mathbf{Z}_0 + \boldsymbol{\Gamma})(\mathbf{I} + (\mathbf{U}_q^\top \mathbf{B}\mathbf{Z}_0)^\dagger \boldsymbol{\Omega})^\top \mathbf{Q}\mathbf{P}^\top (\mathbf{I} + 2\mathrm{sym}(\mathbf{P}\mathbf{Q}^\top \boldsymbol{\Omega}^\top \\ &\quad \mathbf{P}\boldsymbol{\Sigma}^{-1}\mathbf{P}^\top) + \mathbf{P}\boldsymbol{\Sigma}^{-1}\mathbf{P}^\top \boldsymbol{\Omega}\boldsymbol{\Omega}^\top \mathbf{P}\boldsymbol{\Sigma}^{-1}\mathbf{P}^\top)^{-1}\mathbf{P}\boldsymbol{\Sigma}^{-1}\mathbf{P}^\top \begin{bmatrix}p_T^{-1}(\boldsymbol{\Lambda}_k) & \mathbf{0}\end{bmatrix}^\top \| \\ &\le \|p_T^{-1}(\boldsymbol{\Lambda}_k)\|(\|p_T(\boldsymbol{\Lambda}_{-q})\|\|\mathbf{U}_{-q}^\top \mathbf{B}\mathbf{Z}_0\| + \|\boldsymbol{\Gamma}\|)\|\boldsymbol{\Sigma}^{-1}\| \\ &\quad (1 - 2\|\boldsymbol{\Sigma}^{-1}\|\|\boldsymbol{\Omega}\| - (\|\boldsymbol{\Sigma}^{-1}\|\|\boldsymbol{\Omega}\|)^2)^{-1}(1 + \|(\mathbf{U}_q^\top \mathbf{B}\mathbf{Z}_0)^\dagger\|\|\boldsymbol{\Omega}\|). \end{aligned} \tag{20}$$

Assume that

$$\begin{cases} \|\widehat{\boldsymbol{\xi}}_t\|_{\mathbf{B},2} = O(\frac{1}{(T-t+1)T}(\sqrt{\beta})^{t+1}\sin\theta(\mathbf{Z}_0, \mathbf{U}_q)), \\ \|\mathbf{U}_q^\top \mathbf{B}\widehat{\boldsymbol{\xi}}_t\|_2 = O(\frac{1}{(T-t+1)T}(\lambda_q^+)^{t+1}\cos\theta(\mathbf{Z}_0, \mathbf{U}_q)). \end{cases} \tag{21}$$

By Lemma 3.7 and the above assumption, we can bound $\boldsymbol{\Gamma}$ and $\boldsymbol{\Omega}$ in Lemma 3.6 as follows:

$$\begin{aligned} \|\boldsymbol{\Gamma}\| &\le \sum_{t=0}^{T-1} \|q_t(\boldsymbol{\Lambda}_{-q})\|\|\widehat{\boldsymbol{\xi}}_{T-t-1}\|_{\mathbf{B},2} \le \sum_{t=0}^{T-1}(t+1)(\sqrt{\beta})^t \frac{(\sqrt{\beta})^{T-t}\sin\theta(\mathbf{Z}_0, \mathbf{U}_q)}{(t+2)T} \\ &\le (\sqrt{\beta})^T \sin\theta(\mathbf{Z}_0, \mathbf{U}_q), \end{aligned} \tag{22}$$

$$\begin{aligned} \|\boldsymbol{\Omega}\| &\le \sum_{t=0}^{T-1} \|p_T^{-1}(\boldsymbol{\Lambda}_q)q_t(\boldsymbol{\Lambda}_q)\|\|\mathbf{U}_q^\top \mathbf{B}\widehat{\boldsymbol{\xi}}_{T-t-1}\|_2 \le 2\sum_{t=0}^{T-1}(t+1)(\lambda_q^+)^{t-T}\frac{(\lambda_q^+)^{T-t}\cos\theta(\mathbf{Z}_0, \mathbf{U}_q)}{16(t+2)T} \\ &\le \frac{1}{8}\cos\theta(\mathbf{Z}_0, \mathbf{U}_q). \end{aligned} \tag{23}$$

Also note that

$$\|\mathbf{U}_{-q}^\top \mathbf{B}\mathbf{Z}_0\| = \sin\theta(\mathbf{Z}_0, \mathbf{U}_q), \quad \|(\mathbf{U}_q^\top \mathbf{B}\mathbf{Z}_0)^\dagger\| = \|\boldsymbol{\Sigma}^{-1}\| = \cos^{-1}\theta(\mathbf{Z}_0, \mathbf{U}_q). \tag{24}$$

By Lemma 3.7 and Eq. (22)-(24), h_T in Eq. (20) can be further bounded as follows:

$$h_T \le 16(\frac{\sqrt{\beta}}{\lambda_k^+})^T \tan\theta(\mathbf{Z}_0, \mathbf{U}_q).$$

Thus, we get $h_T < \epsilon$ if $T \ge 2\sqrt{\frac{\lambda_k}{\lambda_k - 2\sqrt{\beta}}}\log(\frac{16}{\epsilon}\tan\theta(\mathbf{Z}_0, \mathbf{U}_q))$. Let orthonormal $\mathbf{Z}_T^\perp$ be the orthogonal complement of $\mathbf{Z}_T$ in inner product $\langle , \rangle_{\mathbf{B}}$, i.e., $(\mathbf{Z}_T^\perp)^\top \mathbf{B}\mathbf{Z}_T^\perp = \mathbf{I}$ and $(\mathbf{Z}_T^\perp)^\top \mathbf{B}\mathbf{Z}_T = \mathbf{0}$.

$$\begin{aligned} \sin\theta(\mathbf{Z}_T, \mathbf{U}_k) &= \|(\mathbf{Z}_T^\perp)^\top \mathbf{B}\mathbf{U}_k\|_2 = \|(\mathbf{B}^{\frac{1}{2}}\mathbf{Z}_T^\perp)^\top (\mathbf{B}^{\frac{1}{2}}\mathbf{U}_k)\|_2 \\ &= \|(\mathbf{B}^{\frac{1}{2}}\mathbf{Z}_T^\perp)(\mathbf{B}^{\frac{1}{2}}\mathbf{Z}_T^\perp)^\top (\mathbf{B}^{\frac{1}{2}}\mathbf{U}_k)\|_2 \end{aligned}$$

$$
\begin{aligned}
&= \|(\mathbf{I}-\mathbf{B}^{\frac{1}{2}}\mathbf{Z}_T\mathbf{Z}_T^\top\mathbf{B}^{\frac{1}{2}})(\mathbf{B}^{\frac{1}{2}}\mathbf{U}_k)\|_2 \le \|\mathbf{B}^{\frac{1}{2}}\mathbf{U}_k - \mathbf{B}^{\frac{1}{2}}\mathbf{Z}_T(\mathbf{U}_q^\top\mathbf{B}^{\frac{1}{2}}\mathbf{Z}_T)^\dagger \begin{bmatrix}\mathbf{I}_k & \mathbf{0}\end{bmatrix}^\top\|_2 \\
&= \|\mathbf{U}_q^\top\mathbf{B}^{\frac{1}{2}}(\mathbf{B}^{\frac{1}{2}}\mathbf{U}_k - \mathbf{B}^{\frac{1}{2}}\mathbf{Z}_T(\mathbf{U}_q^\top\mathbf{B}\mathbf{Z}_T)^\dagger \begin{bmatrix}\mathbf{I}_k & \mathbf{0}\end{bmatrix}^\top)\|_2 \\
&\quad + \|\mathbf{U}_{-q}^\top\mathbf{B}^{\frac{1}{2}}(\mathbf{B}^{\frac{1}{2}}\mathbf{U}_k - \mathbf{B}^{\frac{1}{2}}\mathbf{Z}_T(\mathbf{U}_q^\top\mathbf{B}\mathbf{Z}_T)^\dagger \begin{bmatrix}\mathbf{I}_k & \mathbf{0}\end{bmatrix}^\top)\|_2 \\
&= \|\mathbf{U}_{-q}^\top\mathbf{B}\mathbf{Z}_T(\mathbf{U}_q^\top\mathbf{B}\mathbf{Z}_T)^\dagger \begin{bmatrix}\mathbf{I}_k & \mathbf{0}\end{bmatrix}^\top\| = h_T < \epsilon.
\end{aligned}
$$

Note that $\|\widehat{\boldsymbol{\xi}}_t\|_{\mathbf{B},2} \le \|\boldsymbol{\xi}_t\|_{\mathbf{B},2}\|\mathbf{C}_t\|_2$ and $\|\mathbf{U}_q^\top\mathbf{B}\widehat{\boldsymbol{\xi}}_t\|_2 \le \|\mathbf{U}_q^\top\mathbf{B}\boldsymbol{\xi}_t\|_2\|\mathbf{C}_t\|_2$ hold by Eq. (18). When the noise conditions in terms of $\boldsymbol{\xi}_t$ satisfy

$$
\begin{cases}
\|\boldsymbol{\xi}_t\|_{\mathbf{B},2} = O(\frac{1}{(T-t+1)T}(\frac{\sqrt{\beta}}{\lambda_1^+})^t\sqrt{\beta}\sin\theta(\mathbf{Z}_t,\mathbf{U}_q)) \\
\|\mathbf{U}_q^\top\mathbf{B}\boldsymbol{\xi}_t\| = O(\frac{1}{(T-t+1)T}(\frac{\lambda_q^+}{\lambda_1^+})^t\lambda_q^+\cos\theta(\mathbf{Z}_t,\mathbf{U}_q))
\end{cases}, \tag{25}
$$

Assumption 21 will be satisfied by Lemma 3.8 and 3.2.

Note that noise $\boldsymbol{\xi}_t$ is generated from approximating $\mathbf{B}^{-1}\mathbf{A}\mathbf{Z}_t$ by a warm-started least-squares solver. The approximation $\tilde{\mathbf{Z}}_{t+1}$ in Algorithm 2 can be written as

$$\tilde{\mathbf{Z}}_{t+1} = \mathbf{B}^{-1}\mathbf{A}\mathbf{Z}_t + \boldsymbol{\xi}_t.$$

For simplicity, we follow Ge et al. (2016) to use accelerated gradient descent as our least-squares solver for minimizing the following least-squares problem

$$f(\mathbf{Z}) = \mathrm{tr}(\frac{1}{2}\mathbf{Z}^\top\mathbf{B}\mathbf{Z} - \mathbf{Z}^\top\mathbf{B}\mathbf{Z}_t)$$

with warm start $\mathbf{Z}_t\mathbf{H}_t$, where $\mathbf{H}_t = (\mathbf{Z}_t^\top\mathbf{B}\mathbf{Z}_t)^{-1}(\mathbf{Z}_t^\top\mathbf{A}\mathbf{Z}_t)$. It suffices for

$$\|\boldsymbol{\xi}_t\|_{\mathbf{B},2} = O(\frac{1}{(T-t+1)T}\min\{\sqrt{\beta}\sin\theta(\mathbf{Z}_t,\mathbf{U}_q), \lambda_q^+\cos\theta(\mathbf{Z}_t,\mathbf{U}_q)\})$$

to meet the accuracy designated by Eq. (25). To get the complexity of solving this least-squares problem to the above accuracy, we need to figure out both the initial and final errors. By the analysis in Ge et al. (2016), we can have any error $(f(\mathbf{Z}) - \min f(\mathbf{Z}))$ expressed as

$$f(\mathbf{Z}) - \min f(\mathbf{Z}) = \frac{1}{2}\|\mathbf{Z} - \mathbf{B}^{-1}\mathbf{A}\mathbf{Z}_t\|_{\mathbf{B},F}^2$$

and the initial error bounded as $f(\mathbf{Z}_t\mathbf{H}_t) - \min f(\mathbf{Z}) \le 4p\lambda_1^2\tan^2\theta(\mathbf{Z}_t,\mathbf{U}_p)$. The final error is

$$f(\tilde{\mathbf{Z}}_{t+1}) - \min f(\mathbf{Z}) = \frac{1}{2}\|\tilde{\mathbf{Z}}_{t+1} - \mathbf{B}^{-1}\mathbf{A}\mathbf{Z}_t\|_{\mathbf{B},F}^2 = \frac{1}{2}\|\boldsymbol{\xi}_t\|_{\mathbf{B},F}^2 \le \frac{p}{2}\|\boldsymbol{\xi}_t\|_{\mathbf{B},2}^2.$$

Noting that we can write that $\tan\theta(\mathbf{Z}_t,\mathbf{U}_p) \le \tau\tan\theta(\mathbf{Z}_t,\mathbf{U}_q)$ for some positive constant τ, the final to initial error can be bounded as follows

$$
\begin{aligned}
\frac{f(\tilde{\mathbf{Z}}_{t+1}) - \min f(\mathbf{Z})}{f(\mathbf{Z}_t\mathbf{H}_t) - \min f(\mathbf{Z})} &= O(\frac{\lambda_k^2}{\lambda_1^2T^2}\min\{\cos^2\theta(\mathbf{Z}_t,\mathbf{U}_q), \frac{\cos^4\theta(\mathbf{Z}_t,\mathbf{U}_q)}{\sin^2\theta(\mathbf{Z}_t,\mathbf{U}_q)}\}). \\
&= \begin{cases} O((\frac{\rho}{\gamma})^2\cos^4\theta(\mathbf{Z}_0,\mathbf{U}_q)), & \theta(\mathbf{Z}_t,\mathbf{U}_q) \text{ is large} \\ O((\frac{\rho}{\gamma})^2), & \theta(\mathbf{Z}_t,\mathbf{U}_q) \text{ is small} \end{cases} \triangleq O(\delta).
\end{aligned}
$$

Since $\mathcal{T}(\delta) = \mathrm{nnz}(\mathbf{B})\sqrt{\kappa(\mathbf{B})}\log\frac{1}{\delta}$ is the complexity of AGD, we have the total complexity

$$
\begin{aligned}
&\frac{1}{\sqrt{\rho}}\Big(\mathrm{nnz}(\mathbf{A})p + \mathrm{nnz}(\mathbf{B})p + np^2\Big)\log\frac{1}{\epsilon\cos\theta_0} \\
&+\frac{1}{\sqrt{\rho}}\Big(\log\frac{1}{\cos\theta_0}\cdot\mathcal{T}((\frac{\rho}{\gamma})^2\cos^4\theta_0) + \log\frac{1}{\epsilon}\cdot\mathcal{T}((\frac{\rho}{\gamma})^2)\Big) \\
&+\frac{1}{\sqrt{\rho}}\Big(\mathrm{nnz}(\mathbf{B})p^2 + np^2\Big)\log\frac{1}{\epsilon\cos\theta_0},
\end{aligned}
$$

where three parts represent complexities of computing $\mathbf{H}_t$, solving least-squares, and $\mathbf{B}$-orthonormalization, respectively. Plugging the formula of $\mathcal{T}(\delta)$ gives us the simplified complexity in Theorem 4.1. □

Algorithm 3 ANPM for CCA under $p \geq k$.

Input: data matrices $(\mathbf{X}, \mathbf{Y})$, block size k, momentum parameter β, target rank k, iteration rank $p \geq k$, iteration number T, a subroutine $\mathrm{GS}_{\mathbf{B}}$ that performs modified Gram-Schmidt process with inner product $\langle , \rangle_{\mathbf{B}}$.
Output: approximate top-k canonical subspaces spanned by the first k columns of $\mathbf{\Phi}_T$ and $\mathbf{\Psi}_T$, respectively.
Set $d_x \times p$ matrices $\mathbf{\Phi}_{-1} = \mathbf{0}$, $\mathbf{\Phi}_0 = \widetilde{\mathbf{\Phi}}(\widetilde{\mathbf{\Phi}}^\top \mathbf{C}_{xx} \widetilde{\mathbf{\Phi}})^{-\frac{1}{2}}$, and $d_y \times p$ matrices $\mathbf{\Psi}_{-1} = \mathbf{0}$, $\mathbf{\Psi}_0 = \widetilde{\mathbf{\Psi}}(\widetilde{\mathbf{\Psi}}^\top \mathbf{C}_{yy} \widetilde{\mathbf{\Psi}})^{-\frac{1}{2}}$, where $\widetilde{\mathbf{\Phi}}$ and $\widetilde{\mathbf{\Psi}}$ are entry-wise i.i.d. standard normal matrices of size $d_x \times p$ and $d_y \times p$, respectively
for $t = 0, 1, \cdots, T-1$ **do** {**Perform plain alternating least-squares updates**}
  $\widehat{\mathbf{\Phi}}_t \approx \arg\min_{\mathbf{\Phi} \in \mathbb{R}^{d_x \times k}} l_t(\mathbf{\Phi})$ **which starts from the initial** $\mathbf{\Phi}_t(\mathbf{\Phi}_t^\top \mathbf{C}_{xx} \mathbf{\Phi}_t)^{-1}(\mathbf{\Phi}_t^\top \mathbf{C}_{xy} \mathbf{\Psi}_t)$ **to approximately minimize** $l_t(\mathbf{\Phi}) = \frac{1}{2n}\|\mathbf{X}^\top \mathbf{\Phi} - \mathbf{Y}^\top \mathbf{\Psi}_t\|_F^2 + \frac{r_x}{2}\|\mathbf{\Phi}\|_F^2$
  $\widehat{\mathbf{\Psi}}_t \approx \arg\min_{\mathbf{\Psi} \in \mathbb{R}^{d_y \times k}} h_t(\mathbf{\Psi})$ **which starts from the initial** $\mathbf{\Psi}_t(\mathbf{\Psi}_t^\top \mathbf{C}_{yy} \mathbf{\Psi}_t)^{-1}(\mathbf{\Psi}_t^\top \mathbf{C}_{xy}^\top \mathbf{\Phi}_t)$ **to approximately minimize** $h_t(\mathbf{\Psi}) = \frac{1}{2n}\|\mathbf{Y}^\top \mathbf{\Psi} - \mathbf{X}^\top \mathbf{\Phi}_t\|_F^2 + \frac{r_y}{2}\|\mathbf{\Psi}\|_F^2$
  # Perform faster alternating least-squares updates
  $\widehat{\widehat{\mathbf{\Phi}}}_t \approx \arg\min_{\mathbf{\Phi} \in \mathbb{R}^{d_x \times k}} \widehat{l}_t(\mathbf{\Phi})$ **which starts from the initial** $\widehat{\mathbf{\Phi}}_t(\widehat{\mathbf{\Phi}}_t^\top \mathbf{C}_{xx} \widehat{\mathbf{\Phi}}_t)^{-1}(\widehat{\mathbf{\Phi}}_t^\top \mathbf{C}_{xy} \widehat{\mathbf{\Psi}}_t)$ **to approximately minimize** $\widehat{l}_t(\mathbf{\Phi}) = \frac{1}{2n}\|\mathbf{X}^\top \mathbf{\Phi} - \mathbf{Y}^\top \widehat{\mathbf{\Psi}}_t\|_F^2 + \frac{r_x}{2}\|\mathbf{\Phi}\|_F^2$
  if $t = 0$ **then**
    $\mathbf{\Phi}_{t+1} = \mathrm{GS}_{\mathbf{C}_{xx}}(\frac{1}{2}\widehat{\widehat{\mathbf{\Phi}}}_t)$ **such that** $\mathbf{\Phi}_{t+1}\mathbf{R}_{t+1} = \frac{1}{2}\widehat{\widehat{\mathbf{\Phi}}}_t$
  else
    $\mathbf{\Phi}_{t+1} = \mathrm{GS}_{\mathbf{C}_{xx}}(\widehat{\widehat{\mathbf{\Phi}}}_t - \beta \mathbf{\Phi}_{t-1}\mathbf{R}_t^{-1})$ **such that** $(\widehat{\widehat{\mathbf{\Phi}}}_t - \beta \mathbf{\Phi}_{t-1}\mathbf{R}_t^{-1}) = \mathbf{\Phi}_{t+1}\mathbf{R}_{t+1}$
  end if
  $\widehat{\widehat{\mathbf{\Psi}}}_t \approx \arg\min_{\mathbf{\Psi} \in \mathbb{R}^{d_y \times k}} \widehat{h}_t(\mathbf{\Psi})$ **which starts from the initial** $\widehat{\mathbf{\Psi}}_t(\widehat{\mathbf{\Psi}}_t^\top \mathbf{C}_{yy} \widehat{\mathbf{\Psi}}_t)^{-1}(\widehat{\mathbf{\Psi}}_t^\top \mathbf{C}_{xy}^\top \widehat{\mathbf{\Phi}}_t)$ **to approximately minimize** $\widehat{h}_t(\mathbf{\Psi}) = \frac{1}{2n}\|\mathbf{Y}^\top \mathbf{\Psi} - \mathbf{X}^\top \widehat{\mathbf{\Phi}}_t\|_F^2 + \frac{r_y}{2}\|\mathbf{\Psi}\|_F^2$
  if $t = 0$ **then**
    $\mathbf{\Psi}_{t+1} = \mathrm{GS}_{\mathbf{C}_{yy}}(\frac{1}{2}\widehat{\widehat{\mathbf{\Psi}}}_t)$ **such that** $\mathbf{\Psi}_{t+1}\mathbf{S}_{t+1} = \frac{1}{2}\widehat{\widehat{\mathbf{\Psi}}}_t$
  else
    $\mathbf{\Psi}_{t+1} = \mathrm{GS}_{\mathbf{C}_{yy}}(\widehat{\widehat{\mathbf{\Psi}}}_t - \beta \mathbf{\Psi}_{t-1}\mathbf{S}_t^{-1})$ **such that** $(\widehat{\widehat{\mathbf{\Psi}}}_t - \beta \mathbf{\Psi}_{t-1}\mathbf{S}_t^{-1}) = \mathbf{\Psi}_{t+1}\mathbf{S}_{t+1}$
  end if
end for

Theorem 4.2 *Let $k \leq q \leq p$ and assume that $\sigma_q^2 > 2\sqrt{\beta} \geq \sigma_{q+1}^2$. After Algorithm 3 or the update below*

$$\begin{cases} \mathbf{\Phi}_{t+1}\mathbf{R}_{t+1} = \mathbf{C}_{xx}^{-1}\mathbf{C}_{xy}\big(\mathbf{C}_{yy}^{-1}\mathbf{C}_{xy}^\top \mathbf{\Phi}_t + \boldsymbol{\xi}_t^1\big) - \beta \mathbf{\Phi}_{t-1}\mathbf{R}_t^{-1} + \boldsymbol{\xi}_t^2 \in \mathbb{R}^{d_x \times p} \\ \mathbf{\Psi}_{t+1}\mathbf{S}_{t+1} = \mathbf{C}_{yy}^{-1}\mathbf{C}_{xy}^\top\big(\mathbf{C}_{xx}^{-1}\mathbf{C}_{xy}\mathbf{\Psi}_t + \boldsymbol{\eta}_t^1\big) - \beta \mathbf{\Psi}_{t-1}\mathbf{S}_t^{-1} + \boldsymbol{\eta}_t^2 \in \mathbb{R}^{d_y \times p}, \end{cases}$$

runs for $T = O(\frac{1}{\sqrt{\rho}} \log \frac{1}{\epsilon \cos\theta_0})$, we have that $\sin\max\{\theta(\mathbf{\Phi}_T, \mathbf{U}_k), \theta(\mathbf{\Psi}_T, \mathbf{V}_k)\} \leq \epsilon$ in time complexity

$$O(\mathrm{nnz}(\mathbf{X}, \mathbf{Y})p\sqrt{\frac{\kappa(\mathbf{X}, \mathbf{Y})}{\rho}}(\log\frac{1}{\cos\theta_0}\log\frac{\gamma}{\rho\cos\theta_0} + \log\frac{1}{\epsilon}\log\frac{\gamma}{\rho}) + \frac{\mathrm{nnz}(\mathbf{X}, \mathbf{Y})p^2}{\sqrt{\rho}}\log\frac{1}{\epsilon\cos\theta_0}),$$

where $\mathrm{nnz}(\mathbf{X}, \mathbf{Y}) = \mathrm{nnz}(\mathbf{X}) + \mathrm{nnz}(\mathbf{Y})$, $\kappa(\mathbf{X}, \mathbf{Y}) = \max\{\kappa(\mathbf{C}_{xx}), \kappa(\mathbf{C}_{yy})\}$, $\rho = \frac{\sigma_k^2 - 2\sqrt{\beta}}{\sigma_k^2}$, $\gamma = \frac{\sigma_1^2}{\sigma_k^2}$, and $\theta_0 = \max\{\theta(\mathbf{\Phi}_0, \mathbf{U}_q), \theta(\mathbf{\Psi}_0, \mathbf{V}_q)\}$.

Proof The proof works by repeating the proof of Theorem 4.1 twice with $(\mathbf{A}, \mathbf{B}) = (\mathbf{C}_{xy}\mathbf{C}_{yy}^{-1}\mathbf{C}_{xy}^\top, \mathbf{C}_{xx})$ and $(\mathbf{A}, \mathbf{B}) = (\mathbf{C}_{xy}^\top\mathbf{C}_{xx}^{-1}\mathbf{C}_{xy}, \mathbf{C}_{yy})$, respectively, except for the handling of noise terms because of two sources of noise for each step. We take as an example the case of $(\mathbf{A}, \mathbf{B}) = (\mathbf{C}_{xy}\mathbf{C}_{yy}^{-1}\mathbf{C}_{xy}^\top, \mathbf{C}_{xx})$, where it holds that

$$\mathbf{C}_{xy} = \mathbf{C}_{xx}\left(\mathbf{U}_j\mathbf{\Sigma}_j\mathbf{V}_j^\top + \mathbf{U}_{-j}\mathbf{\Sigma}_{-j}\mathbf{V}_{-j}^\top\right)\mathbf{C}_{yy},$$

and

$$\mathbf{C}_{xy}\mathbf{C}_{yy}^{-1}\mathbf{C}_{xy}^{\top} = \mathbf{C}_{xx}\left(\mathbf{U}_j\mathbf{\Sigma}_j^2\mathbf{U}_j^{\top} + \mathbf{U}_{-j}\mathbf{\Sigma}_{-j}^2\mathbf{U}_{-j}^{\top}\right)\mathbf{C}_{xx}.$$

In this case, the noise for each step is $(\mathbf{C}_{xx}^{-1}\mathbf{C}_{xy}\boldsymbol{\xi}_t^1 + \boldsymbol{\xi}_t^2)$, due to the two least-squares approximations, i.e.,

$$\widehat{\mathbf{\Psi}}_t = \mathbf{C}_{yy}^{-1}\mathbf{C}_{xy}^{\top}\mathbf{\Phi}_t + \boldsymbol{\xi}_t^1, \qquad \widehat{\widehat{\mathbf{\Phi}}}_t = \mathbf{C}_{xx}^{-1}\mathbf{C}_{xy}\widehat{\mathbf{\Psi}}_t + \boldsymbol{\xi}_t^2.$$

It suffices to assume that the final approximation accuracies for two least-squares problems in Lines 5 and 7 of Algorithm 3 satisfy

$$\begin{aligned} h_t(\widehat{\mathbf{\Psi}}_t) - \min h_t(\mathbf{\Psi}) &= \frac{1}{2}\|\widehat{\mathbf{\Psi}}_t - \mathbf{C}_{yy}^{-1}\mathbf{C}_{xy}^{\top}\mathbf{\Phi}_t\|_{\mathbf{C}_{yy},F}^2 = \frac{1}{2}\|\boldsymbol{\xi}_t^1\|_{\mathbf{C}_{yy},F}^2 \leq \frac{p}{2}\|\boldsymbol{\xi}_t^1\|_{\mathbf{C}_{yy},2}^2, \\ \widehat{l}_t(\widehat{\widehat{\mathbf{\Phi}}}_t) - \min \widehat{l}_t(\mathbf{\Phi}) &= \frac{1}{2}\|\widehat{\widehat{\mathbf{\Phi}}}_t - \mathbf{C}_{xx}^{-1}\mathbf{C}_{xy}\widehat{\mathbf{\Psi}}_t\|_{\mathbf{C}_{xx},F}^2 = \frac{1}{2}\|\boldsymbol{\xi}_t^2\|_{\mathbf{C}_{xx},F}^2 \leq \frac{p}{2}\|\boldsymbol{\xi}_t^2\|_{\mathbf{C}_{xx},2}^2, \end{aligned}$$

where

$$\begin{aligned} \|\boldsymbol{\xi}_t^1\|_{\mathbf{C}_{yy},2} &= O(\frac{1}{(T-t+1)T}\min\{\sqrt{\beta}\sin\theta_t, (\sigma_q^2)^+\cos\theta_t\}), \\ \|\boldsymbol{\xi}_t^2\|_{\mathbf{C}_{xx},2} &= O(\frac{1}{(T-t+1)T}\min\{\sqrt{\beta}\sin\theta_t, (\sigma_q^2)^+\cos\theta_t\}), \end{aligned}$$

$\theta_t = \max\{\theta(\mathbf{\Phi}_t, \mathbf{U}_q), \theta(\mathbf{\Psi}_t, \mathbf{V}_q)\}$, and $(\sigma_j^2)^{\pm} = \frac{\sigma_j^2 \pm \sqrt{\sigma_j^4 - 4\beta}}{2}$. By Lemmas 3.2-3.3 in Xu and Li (2021), we have the following initial errors

$$\begin{aligned} h_t(\mathbf{\Psi}_t\mathbf{H}_{\mathbf{\Psi}_t}) - \min h_t(\mathbf{\Psi}) &= \frac{1}{2}\|\mathbf{\Psi}_t\mathbf{H}_{\mathbf{\Psi}_t} - \mathbf{C}_{yy}^{-1}\mathbf{C}_{xy}^{\top}\mathbf{\Phi}_t\|_{\mathbf{C}_{yy},F}^2 = O(p\sigma_1^2\tan^2\theta_t), \\ \widehat{l}_t(\widehat{\mathbf{\Phi}}_t\mathbf{H}_{\widehat{\mathbf{\Phi}}_t}) - \min \widehat{l}_t(\mathbf{\Phi}) &= \frac{1}{2}\|\widehat{\mathbf{\Phi}}_t\mathbf{H}_{\widehat{\mathbf{\Phi}}_t} - \mathbf{C}_{xx}^{-1}\mathbf{C}_{xy}\widehat{\mathbf{\Psi}}_t\|_{\mathbf{C}_{xx},F}^2 = O(p(\sigma_1^2 + \|\boldsymbol{\xi}_t^1\|_{\mathbf{C}_{yy},2}^2)\tan^2\theta_t), \end{aligned}$$

where $\mathbf{H}_{\mathbf{\Psi}_t} = (\mathbf{\Psi}_t^{\top}\mathbf{C}_{yy}\mathbf{\Psi}_t)^{-1}(\mathbf{\Psi}_t^{\top}\mathbf{C}_{xy}^{\top}\mathbf{\Phi}_t)$ and $\mathbf{H}_{\widehat{\mathbf{\Phi}}_t} = (\widehat{\mathbf{\Phi}}_t^{\top}\mathbf{C}_{xx}\widehat{\mathbf{\Phi}}_t)^{-1}(\widehat{\mathbf{\Phi}}_t^{\top}\mathbf{C}_{xy}\widehat{\mathbf{\Psi}}_t)$. Thus, we can bound the final to initial error ratio as follows

$$\begin{aligned} \frac{h_t(\widehat{\mathbf{\Psi}}_t) - \min h_t(\mathbf{\Psi})}{h_t(\mathbf{\Psi}_t\mathbf{H}_{\mathbf{\Psi}_t}) - \min h_t(\mathbf{\Psi})} &= O(\frac{\sigma_k^2}{\sigma_1^2T^2}\min\{\cos^2\theta_t, \frac{\cos^4\theta_t}{\sin^2\theta_t}\}), \\ \frac{\widehat{l}_t(\widehat{\widehat{\mathbf{\Phi}}}_t) - \min \widehat{l}_t(\mathbf{\Phi})}{\widehat{l}_t(\widehat{\mathbf{\Phi}}_t\mathbf{H}_{\widehat{\mathbf{\Phi}}_t}) - \min \widehat{l}_t(\mathbf{\Phi})} &= O(\frac{\sigma_k^2}{\sigma_1^2T^2}\min\{\cos^2\theta_t, \frac{\cos^4\theta_t}{\sin^2\theta_t}\}). \end{aligned}$$

We then can similarly have the following complexity

$$O(\mathrm{nnz}(\mathbf{X},\mathbf{Y})p\sqrt{\frac{\kappa(\mathbf{X},\mathbf{Y})}{\rho}}(\log\frac{1}{\cos\theta_0}\log\frac{\gamma}{\rho\cos\theta_0} + \log\frac{1}{\epsilon}\log\frac{\gamma}{\rho}) + \frac{\mathrm{nnz}(\mathbf{X},\mathbf{Y})p^2}{\sqrt{\rho}}\log\frac{1}{\epsilon\cos\theta_0}).$$

□

C MOMENTUM PARAMETER IN OTHER RANGES

For ease of exposition, we consider the case of Section 3.1 of the main text, but it is straightforward to extend to the generalized eigenspace computation. In this case, we can resume the proof in Section 3.1 of the main text from Eq. (20) for each case of $\beta > 0$.

1) When $2\sqrt{\beta} \geq \lambda_k$, Algorithm 1 is not guaranteed to converge. The reason is as follows.

Proof In this case, we have that

$$\|p_T^{-1}(\mathbf{\Sigma}_k)\| = \frac{2}{(\sqrt{\beta})^T} \max_{m:2\sqrt{\beta}\geq\lambda_m\geq\lambda_k} |\cos(T\arccos(\frac{\lambda_m}{2\sqrt{\beta}}))|^{-1}, \quad \|p_T(\mathbf{\Sigma}_{-q})\| \leq (\sqrt{\beta})^T,$$

where we have used the equivalent expression for $p_t(x)$ when $|x| < 4\sqrt{\beta}$. Even if the noise is well conditioned such that all the relevant factors are properly bounded in Eq. (20) as in the case considered in the main text, we can only get that

$$\begin{aligned} h_T &\leq c(\sqrt{\beta})^T \frac{2}{(\sqrt{\beta})^T} \max_{m:2\sqrt{\beta}\geq\lambda_m\geq\lambda_k} |\cos(T\arccos(\frac{\lambda_m}{2\sqrt{\beta}}))|^{-1} \\ &\leq 2c \max_{m:2\sqrt{\beta}\geq\lambda_m\geq\lambda_k} |\cos(T\arccos(\frac{\lambda_m}{2\sqrt{\beta}}))|^{-1}, \end{aligned}$$

where c is a constant. In this case, there is no guarantee that $h_T = O(\epsilon)$ can be achieved. Thus, $\sin\theta(\mathbf{X}_T, \mathbf{U}_k) < \epsilon$ is not guaranteed since $\sin\theta(\mathbf{X}_T, \mathbf{U}_k) < h_T$. □

2) When $\lambda_k > 2\sqrt{\beta} \geq \lambda_q$, we have $\sin\theta(\mathbf{X}_T, \mathbf{U}_k) < \epsilon$ for

$$T \geq 2\sqrt{\frac{\lambda_k}{\lambda_k - 2\sqrt{\beta}}} \log(\frac{16}{\epsilon}\tan\theta(\mathbf{X}_0, \mathbf{U}_q))$$

if it holds that

$$\begin{cases} \|\boldsymbol{\xi}_t\| = O(\frac{1}{(T-t+1)T}(\frac{\sqrt{\beta}}{\lambda_1^+})^{t+1}\sqrt{\beta}\sin\theta(\mathbf{X}_0, \mathbf{U}_q)) \\ \|\mathbf{U}_q^\top\boldsymbol{\xi}_t\| = O(\frac{1}{(T-t+1)T}(\frac{\lambda_q^+}{\lambda_1^+})^{t+1}\lambda_q^+\cos\theta(\mathbf{X}_0, \mathbf{U}_q)) \end{cases}.$$

Proof In this case, results in Lemma 3.7 become as follows.

$$\|p_T^{-1}(\mathbf{\Sigma}_k)\| \leq \frac{2}{(\lambda_k^+)^T}, \quad \|p_T(\mathbf{\Sigma}_{-q})\| \leq (\sqrt{\beta})^T, \quad \|q_t(\mathbf{\Sigma}_{-q})\| \leq (t+1)(\sqrt{\beta})^t,$$

$$\begin{aligned} \|p_T^{-1}(\mathbf{\Sigma}_q)q_t(\mathbf{\Sigma}_q)\| &= \max_{1\leq m\leq q} |\frac{\sum_{j=0}^t (\lambda_m^+)^{t-j}(\lambda_m^-)^j}{\frac{1}{2}((\lambda_m^+)^T + (\lambda_m^-)^T)}| \\ = \max\{ &\max_{m:\lambda_m\geq 2\sqrt{\beta}} \frac{\sum_{j=0}^t (\lambda_m^+)^{t-j}(\lambda_m^-)^j}{\frac{1}{2}((\lambda_m^+)^T + (\lambda_m^-)^T)}, \max_{m:2\sqrt{\beta}>\lambda_m\geq\lambda_q} |\frac{\sum_{j=0}^t (\lambda_m^+)^{t-j}(\lambda_m^-)^j}{\frac{1}{2}((\lambda_m^+)^T + (\lambda_m^-)^T)}|\} \\ \leq \max\{ &\max_{m:\lambda_m\geq 2\sqrt{\beta}} \frac{2(t+1)(\lambda_m^+)^t}{(\lambda_m^+)^T}, \max_{m:2\sqrt{\beta}>\lambda_m\geq\lambda_q} \frac{\sum_{j=0}^t |\lambda_m^+|^{t-j}|\lambda_m^-|^j}{|\frac{1}{2}((\lambda_m^+)^T + (\lambda_m^-)^T)|}\} \\ = \max\{ &\max_{m:\lambda_m\geq 2\sqrt{\beta}} \frac{2(t+1)}{(\lambda_m^+)^{T-t}}, \max_{m:2\sqrt{\beta}>\lambda_m\geq\lambda_q} \frac{\sum_{j=0}^t (\sqrt{\beta})^t}{(\sqrt{\beta})^T|\cos(T\arccos\frac{\lambda_m}{2\sqrt{\beta}})|}\} \\ \leq \max\{ &\frac{2(t+1)}{(\sqrt{\beta})^{T-t}}, \max_{m:2\sqrt{\beta}>\lambda_m\geq\lambda_q} \frac{t+1}{(\sqrt{\beta})^{T-t}|\cos(T\arccos\frac{\lambda_m}{2\sqrt{\beta}})|}\} \\ \leq (t+1)(\sqrt{\beta})^{t-T} &\underbrace{\max\{2, \max_{m:2\sqrt{\beta}>\lambda_m\geq\lambda_q} |\cos(T\arccos\frac{\lambda_m}{2\sqrt{\beta}})|^{-1}\}}_{\tau}. \end{aligned}$$

We now assume that

$$\begin{cases} \|\widehat{\boldsymbol{\xi}}_t\| = O(\frac{1}{(T-t+1)T}(\sqrt{\beta})^{t+1}\sin\theta(\mathbf{X}_0, \mathbf{U}_q)), \\ \|\mathbf{U}_q^\top\widehat{\boldsymbol{\xi}}_t\| = O(\frac{1}{\tau(T-t+1)T}(\sqrt{\beta})^{t+1}\cos\theta(\mathbf{X}_0, \mathbf{U}_q)), \end{cases}$$

and can bound $\mathbf{\Gamma}$ and $\mathbf{\Omega}$ in Lemma 3.6 as follows.

$$\begin{aligned} \|\mathbf{\Gamma}\| &\leq \sum_{t=0}^{T-1} \|q_t(\mathbf{\Sigma}_{-q})\|\|\widehat{\boldsymbol{\xi}}_{T-t-1}\| \\ &\leq \sum_{t=0}^{T-1}(t+1)(\sqrt{\beta})^t \cdot \frac{(\sqrt{\beta})^{T-t}\sin\theta(\mathbf{X}_0,\mathbf{U}_q)}{(t+2)T} \leq (\sqrt{\beta})^T\sin\theta(\mathbf{X}_0, \mathbf{U}_q). \end{aligned}$$

$$\begin{aligned}\|\mathbf{\Omega}\| &\leq \textstyle\sum_{t=0}^{T-1} \|p_T^{-1}(\mathbf{\Sigma}_q) q_t(\mathbf{\Sigma}_q)\| \|\mathbf{U}_q^\top \widehat{\boldsymbol{\xi}}_{T-t-1}\| \\ &\leq 2\textstyle\sum_{t=0}^{T-1} \tau(t+1)(\sqrt{\beta})^{t-T} \cdot \frac{(\sqrt{\beta})^{T-t}\cos\theta(\mathbf{X}_0,\mathbf{U}_q)}{16\tau(t+2)T} \leq (1/8)\cos\theta(\mathbf{X}_0,\mathbf{U}_q).\end{aligned}$$

Thus, we can have h_T in Eq. (20) bounded as

$$h_T \leq 16(\frac{\sqrt{\beta}}{\lambda_k^+})^T \tan\theta(\mathbf{X}_0,\mathbf{U}_q)$$

again. Similarly, we have $\sin\theta(\mathbf{X}_T,\mathbf{U}_k) \leq h_T < \epsilon$ for

$$T \geq 2\sqrt{\frac{\lambda_k}{\lambda_k - 2\sqrt{\beta}}}\log(\frac{16}{\epsilon}\tan\theta(\mathbf{X}_0,\mathbf{U}_q)).$$

To convert noise conditions, noting $\tau \geq 2$ we have that

$$\begin{aligned}\|\mathbf{C}_t\| &\leq (\lambda_1^+)^t + \tfrac{1}{16}\textstyle\sum_{j=0}^{t-1} \frac{(\sqrt{\beta})^{t-j}\cos\theta(\mathbf{X}_0,\mathbf{U}_q)}{\tau(T-t+j+2)T}(j+1)(\lambda_1^+)^j \\ &\leq (\lambda_1^+)^t + \frac{(\lambda_1^+)^t\cos\theta(\mathbf{X}_0,\mathbf{U}_q)}{16T}\textstyle\sum_{j=0}^{t-1}\frac{j+1}{T-t+j+2} \leq 2(\lambda_1^+)^t,\end{aligned}$$

and $\|\mathbf{C}_t\| \geq \frac{7}{16}(\lambda_1^+)^t$ similarly. Thus, we can write the final noise condition as

$$\begin{cases}\|\boldsymbol{\xi}_t\| = O(\frac{1}{(T-t+1)T}(\frac{\sqrt{\beta}}{\lambda_1^+})^t\sqrt{\beta}\sin\theta(\mathbf{X}_0,\mathbf{U}_q)) \\ \|\mathbf{U}_q^\top\boldsymbol{\xi}_t\| = O(\frac{1}{\tau(T-t+1)T}(\frac{\sqrt{\beta}}{\lambda_1^+})^t\sqrt{\beta}\cos\theta(\mathbf{X}_0,\mathbf{U}_q))\end{cases}.$$

□

3) When $\lambda_{q+1} > 2\sqrt{\beta}$, we have $\sin\theta(\mathbf{X}_T,\mathbf{U}_k) < \epsilon$ for

$$T \geq \frac{\lambda_k^+}{\lambda_k^+ - \lambda_{q+1}^+}\log(\frac{16}{\epsilon}\tan\theta(\mathbf{X}_0,\mathbf{U}_q))$$

if

$$\begin{cases}\|\boldsymbol{\xi}_t\| = O(\frac{1}{(T-t+1)T}(\frac{\lambda_{q+1}^+}{\lambda_1^+})^t\lambda_{q+1}^+\sin\theta(\mathbf{X}_0,\mathbf{U}_q)) \\ \|\mathbf{U}_q^\top\boldsymbol{\xi}_t\| = O(\frac{1}{2(T-t+1)T}(\frac{\lambda_q^+}{\lambda_1^+})^t\lambda_q^+\cos\theta(\mathbf{X}_0,\mathbf{U}_q))\end{cases}.$$

Proof In this case, results in Lemma 3.7 become as follows.

$$\begin{aligned}\|p_T^{-1}(\mathbf{\Sigma}_k)\| &\leq \frac{2}{(\lambda_k^+)^T}, \\ \|p_T(\mathbf{\Sigma}_{-q})\| &\leq \max_{q+1\leq m\leq n}\frac{1}{2}|(\lambda_m^+)^T + (\lambda_m^-)^T| \\ &= \max\{\max_{m:\lambda_m\geq 2\sqrt{\beta}}\tfrac{1}{2}((\lambda_m^+)^T + (\lambda_m^-)^T), \max_{m:2\sqrt{\beta}>\lambda_m}\frac{1}{2}|(\lambda_m^+)^T + (\lambda_m^-)^T|\} \\ &\leq \max\{(\lambda_{q+1}^+)^T, (\sqrt{\beta})^T\} = (\lambda_{q+1}^+)^T, \\ \|q_t(\mathbf{\Sigma}_{-q})\| &\leq \max_{q+1\leq m\leq n}\sum_{j=0}^t|(\lambda_m^+)^{t-j}(\lambda_m^-)^j| \\ &= \max\{\max_{m:\lambda_m\geq 2\sqrt{\beta}}\sum_{j=0}^t(\lambda_m^+)^{t-j}(\lambda_m^-)^j, \max_{m:2\sqrt{\beta}>\lambda_m}|\sum_{j=0}^t(\lambda_m^+)^{t-j}(\lambda_m^-)^j|\} \\ &\leq \max\{(t+1)(\lambda_{q+1}^+)^T, (t+1)(\sqrt{\beta})^T\} = (t+1)(\lambda_{q+1}^+)^T, \\ \|p_T^{-1}(\mathbf{\Sigma}_q)q_t(\mathbf{\Sigma}_q)\| &= \max_{1\leq m\leq q}\frac{\sum_{j=0}^t(\lambda_m^+)^{t-j}(\lambda_m^-)^j}{\frac{1}{2}((\lambda_m^+)^T + (\lambda_m^-)^T)} \leq 2\max_{1\leq m\leq q}\frac{\sum_{j=0}^t(\lambda_m^+)^t}{(\lambda_m^+)^T} = (t+1)(\lambda_q^+)^{t-T}.\end{aligned}$$

We now assume that

$$\begin{cases} \|\widehat{\boldsymbol{\xi}}_t\| = O(\frac{1}{(T-t+1)T}(\lambda_{q+1}^+)^{t+1}\sin\theta(\mathbf{X}_0,\mathbf{U}_q)), \\ \|\mathbf{U}_q^\top\widehat{\boldsymbol{\xi}}_t\| = O(\frac{1}{(T-t+1)T}(\lambda_q^+)^{t+1}\cos\theta(\mathbf{X}_0,\mathbf{U}_q)), \end{cases}$$

and can bound $\boldsymbol{\Gamma}$ and $\boldsymbol{\Omega}$ in Lemma 3.6 as follows.

$$\begin{aligned} \|\boldsymbol{\Gamma}\| &\leq \textstyle\sum_{t=0}^{T-1}\|q_t(\boldsymbol{\Sigma}_{-q})\|\|\widehat{\boldsymbol{\xi}}_{T-t-1}\| \\ &\leq \textstyle\sum_{t=0}^{T-1}(t+1)(\lambda_{q+1}^+)^t\cdot\frac{(\lambda_{q+1}^+)^{T-t}\sin\theta(\mathbf{X}_0,\mathbf{U}_q)}{(t+2)T} \leq (\lambda_{q+1}^+)^T\sin\theta(\mathbf{X}_0,\mathbf{U}_q). \end{aligned}$$

$$\begin{aligned} \|\boldsymbol{\Omega}\| &\leq \textstyle\sum_{t=0}^{T-1}\|p_T^{-1}(\boldsymbol{\Sigma}_q)q_t(\boldsymbol{\Sigma}_q)\|\|\mathbf{U}_q^\top\widehat{\boldsymbol{\xi}}_{T-t-1}\| \\ &\leq 2\textstyle\sum_{t=0}^{T-1}(t+1)(\lambda_q^+)^{t-T}\cdot\frac{(\lambda_q^+)^{T-t}\cos\theta(\mathbf{X}_0,\mathbf{U}_q)}{16\tau(t+2)T} \leq (1/8)\cos\theta(\mathbf{X}_0,\mathbf{U}_q). \end{aligned}$$

Thus, we can have h_T in Eq. (20) bounded as

$$h_T \leq 16(\frac{\lambda_{q+1}^+}{\lambda_k^+})^T\tan\theta(\mathbf{X}_0,\mathbf{U}_q).$$

When $T > \log^{-1}(\lambda_k^+/\lambda_{q+1}^+)\log(16\tan\theta(\mathbf{X}_0,\mathbf{U}_q)/\epsilon)$, we have $h_T < \epsilon$. Noting that

$$\begin{aligned} \log\frac{\lambda_k^+}{\lambda_{q+1}^+} \geq& \frac{-1+\lambda_k^+/\lambda_{q+1}^+}{\lambda_k^+/\lambda_{q+1}^+} = \frac{\lambda_k^+-\lambda_{q+1}^+}{\lambda_k^+} \\ =& \frac{\lambda_k-\lambda_{q+1}+\sqrt{\lambda_k^2-4\beta}-\sqrt{\lambda_{q+1}^2-4\beta}}{\lambda_k+\sqrt{\lambda_k^2-4\beta}} > \frac{\lambda_k-\lambda_{q+1}}{\lambda_k}, \end{aligned}$$

we get $\sin\theta(\mathbf{X}_T,\mathbf{U}_k) \leq h_T < \epsilon$ for $T \geq \frac{\lambda_k^+}{\lambda_k^+-\lambda_{q+1}^+}\log(\frac{16}{\epsilon}\tan\theta(\mathbf{X}_0,\mathbf{U}_q))$. To convert noise conditions, we have that

$$\begin{aligned} \|\mathbf{C}_t\| &\leq (\lambda_1^+)^t + \tfrac{1}{16}\textstyle\sum_{j=0}^{t-1}\frac{(\lambda_q^+)^{t-j}\cos\theta(\mathbf{X}_0,\mathbf{U}_q)}{(T-t+j+2)T}(j+1)(\lambda_1^+)^j \\ &\leq (\lambda_1^+)^t + \frac{(\lambda_1^+)^t\cos\theta(\mathbf{X}_0,\mathbf{U}_q)}{16T}\textstyle\sum_{j=0}^{t-1}\frac{j+1}{T-t+j+2} \leq 2(\lambda_1^+)^t, \end{aligned}$$

and $\|\mathbf{C}_t\| \geq \frac{7}{16}(\lambda_1^+)^t$ similarly. Thus, the final noise condition is

$$\begin{cases} \|\boldsymbol{\xi}_t\| = O(\frac{1}{(T-t+1)T}(\frac{\lambda_{q+1}^+}{\lambda_1^+})^t\lambda_{q+1}^+\sin\theta(\mathbf{X}_0,\mathbf{U}_q)) \\ \|\mathbf{U}_q^\top\boldsymbol{\xi}_t\| = O(\frac{1}{(T-t+1)T}(\frac{\lambda_q^+}{\lambda_1^+})^t\lambda_q^+\cos\theta(\mathbf{X}_0,\mathbf{U}_q)) \end{cases}.$$

□

D MORE EXPERIMENTS

More experiments are provided here to cover more settings.

D.1 ANPM

Two noise settings are considered. First, noise variance σ during iterations is considered. Three values of fixed variance are used: $\sigma \in \{10^{-10}, 10^{-8}, 10^{-5}\}$. The performance of the algorithms is reported in Figure 4, where we can see that the ANPM with $p > k$ runs fastest across datasets for each noise variance value. Again, the two β settings of the ANPM perform almost equally well. We also observe that the iteration error at convergence is positively correlated with, even

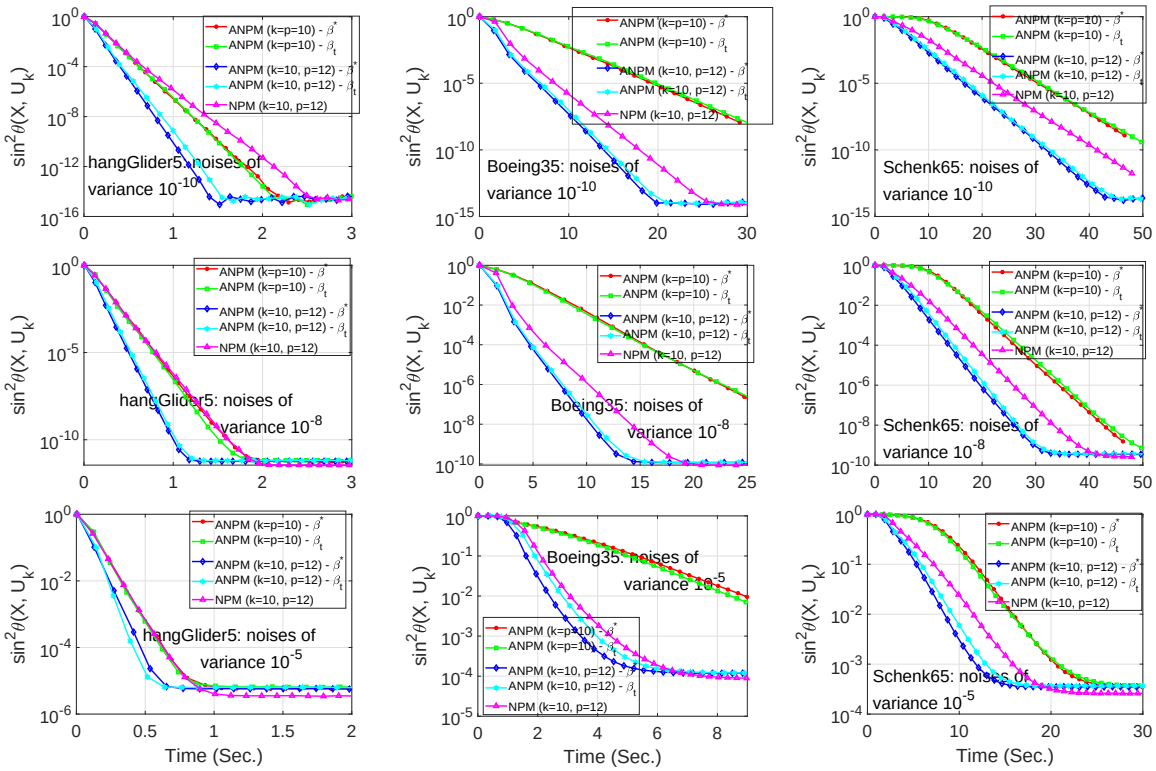

Figure 4: Performance of the ANPM (Algorithm 1) under noises of fixed variance across iterations.

roughly matching, the noise level σ for both types of algorithm, that is, more noises lead to an output solution with a larger error. NPM with $p > k$ works better than ANPM with $p = k$ on the first dataset, but is worse on the other two datasets.

Second, as in the main text, we test the algorithms with noises of varying variance injected into iterations for two more initial noise variances: $\sigma_0 \in \{10, 10^3\}$. Again, the noise variance keeps decreasing with iterations as follows: $\sigma_t = \frac{\sigma_0}{1.1^t}$. The convergence curves of the algorithms are plotted in Figure 5. We can also see the ANPM with $p > k$ performs best.

Next, we check the performance of the ANPM with $p > k$ for different values of p in two noise variance settings (fixed or dynamic noise variance). Figures 6-7 show that increasing p does not always mean better performance. It depends on three factors, i.e., gap $\frac{\lambda_k - \lambda_{p+1}}{\lambda_k}$, iteration rank p, and noise type, where performance increases with the first factor but decreases with the second one.

D.2 ANPM for Genearlized Eigenspace Computation

We also check the performance of the ANPM with $p > k$ for generalized eigenspace computation for different values of p, report the algorithm performance in Figure 8, where we have similar observations as with Figures 6-7.

D.3 ANPM for CCA

For the case of CCA, we test other values of p for the ANPM with $p > k$ and report the performance in Figure 9.

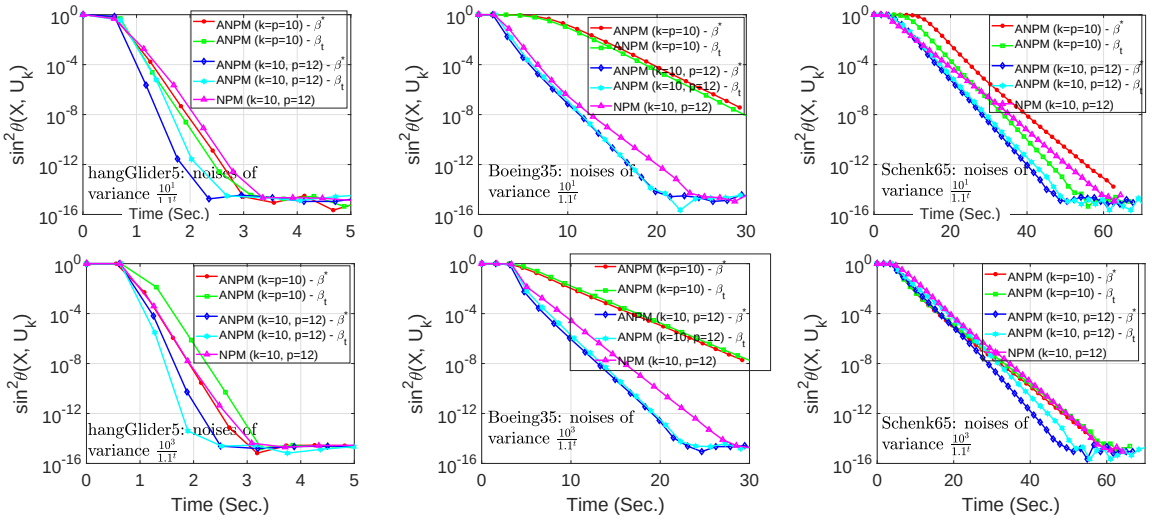

Figure 5: Performance of the ANPM (Algorithm 1) under noises of varying variance.

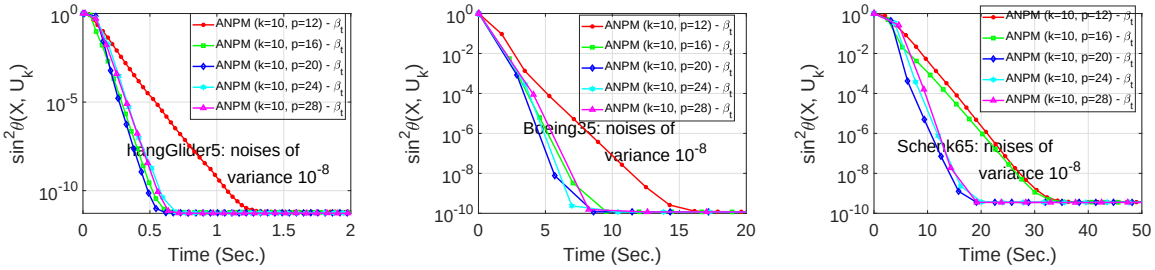

Figure 6: Performance of the ANPM (Algorithm 1) with varying p under noises of fixed variance across iterations.

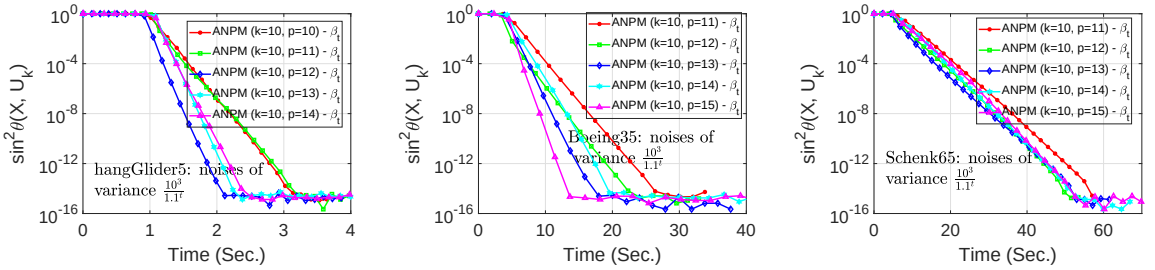

Figure 7: Performance of the ANPM (Algorithm 1) with varying p under noises of varying variance.

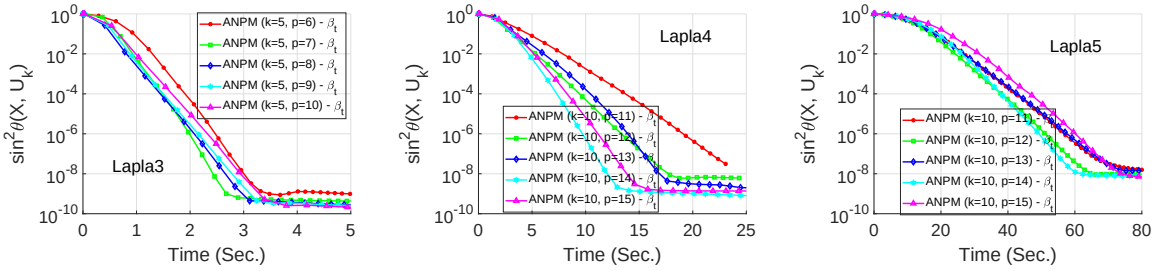

Figure 8: Performance of the ANPM (Algorithm 2) with varying p for generalized eigenspace computation.

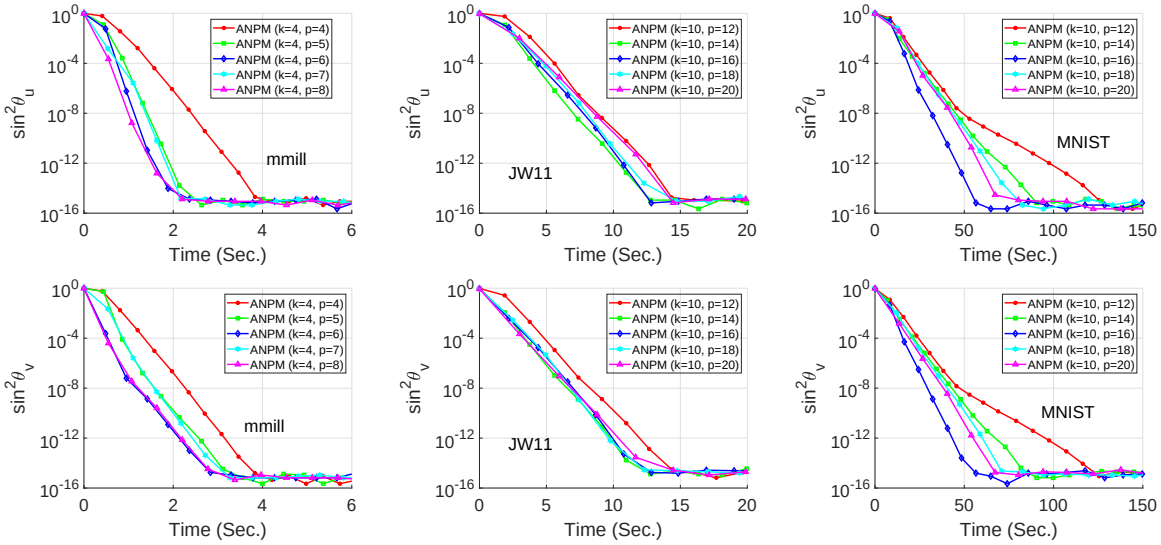

Figure 9: Performance of the ANPM (Algorithm 3) with varying p for CCA.