Some Insights into the Power Method

Pravin Singh, Shivani Singh and Virath Singh

***Abstract*—In this paper, the power method is discussed in mathematical detail with respect to the convergence, analysis, numerical computations and multiplicity. Deflation as well as avoiding deflation are described for symmetric positive definite matrices. Variations of the power method are discussed. Projections are applied to some distributions of the spectrum.**

***Index Terms*—Power method, eigenvalues, projections, deflation.**

I. Introduction

THE spectrum of a matrix determines its eigendecomposition which reveals much about the associated linear transformation. Methods such as those due to Lanczos, Arnoldi and Leverrier, amongst others, have been studied extensively to obtain the spectrum [3]. Amongst all methods, the QR algorithm is the gold standard to determine the spectrum. A pioneering classic text on the computation of eigenvalues is by Wilkinson [12]. However, there are cases where only the dominant and least dominant eigenvalues are needed. For example, the condition of a symmetric linear system is determined by $\left|\frac{\lambda_1}{\lambda_N}\right|$, where λ_1 is the dominant eigenvalue and λ_N the least dominant of the matrix associated with the linear system. The power method is a simple yet effective method to compute λ_1 and λ_N. However, with some simple adaptation, it can be used to determine other eigenvalues as well. The convergence rate can be speeded up, if the need arises. However, there is much more to the power method, that we examine in this treatise. The PageRank is calculated using the power method to determine the principal eigenvector of the Google matrix [6]. The power iteration is still used as part of more efficient techniques like Krylov methods and the QR method. Householder attributed the power method to Müntz in 1913 [4]. A parametric power method has recently been proposed in [1] and is promising, but only for certain distributions of the spectrum. The power iteration using $(\cdot)^{2^k}$ has been shown to be faster than the traditional power method in [8], as the rate of convergence depends on $\left|\frac{\lambda_2}{\lambda_1}\right|^{2^k}$ and is especially useful when $|\frac{\lambda_2}{\lambda_1}|$ is close to unity, for not very large matrix dimensions. A modified power method has been applied in Nuclear Physics and is still a subject of current research [9]. Bounds on the eigenvalues of preconditioned matrices are invaluable as illustrated in [13]. The authors in [11], surprisingly used the power and the inverse power method to determine outer bounds for irreducible positive definite matrices.

Manuscript received September 26, 2022; revised March 12, 2023.

P. Singh is a professor in the Department of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal, Private Bag X54001, Durban, KZN , 4001, South Africa (e-mail: singhp@ukzn.ac.za).

S. Singh is a lecturer in the Department of Decision Science, University of South Africa, PO Box 392, Pretoria, Gauteng, 0003, South Africa (e-mail: singhs2@unisa.ac.za).

V. Singh is a senior lecturer in the Department of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal, Private Bag X54001, Durban, KZN , 4001, South Africa (corresponding author phone:+27 031 2607687; fax: +27 031 2607806; e-mail: singhv@ukzn.ac.za).

II. Theory

It is well known that Hermitian matrices are unitarily diagonalizable and that their eigenvalues are real. Let $\mathbf{A} \in \mathbb{R}^{n\times n}$ be Hermitian with $N \leq n$ distinct eigenvalues $\{\lambda_i\}_{i=1}^{N}$. We denote the algebraic multiplicity of λ_i by m_i and assume that they satisfy the descending arrangement

$$|\lambda_1| > |\lambda_2| > |\lambda_3| > \cdots > |\lambda_{N-1}| > |\lambda_N|.$$

The corresponding orthonormal eigenbasis of $\mathbf{A}$ is given by

$$S = \{\mathbf{u}_i^j \,|i = 1,\, 2, \cdots,\, N; j = 1,\, 2, \cdots,\, m_i\}.$$

We shall refer to λ_i as **dominant** to λ_j, whenever $|\lambda_i| > |\lambda_j|$. We call the tuple $(\lambda_i, \mathbf{u}_i^j)$ the $(i,j)_{th}$ eigenmode. Sometimes we omit the superscript j if we are focussing on one eigenvector and simply refer to the mode as the i_{th} mode. Define the subspaces

$$S_j = \bigoplus_{i=1}^{j} N(\mathbf{A} - \lambda_i \mathbf{I}) \text{ and } \hat{S}_j = \bigoplus_{i=j}^{N} N(\mathbf{A} - \lambda_i \mathbf{I}),$$

where $N(\mathbf{A} - \lambda_i \mathbf{I})$ denotes the nullspace of $\mathbf{A} - \lambda_i \mathbf{I}$ with dimension m_i.

Theorem 1: Let $\mathbf{A} \in \mathbb{R}^{n\times n}$ be Hermitian with spectrum, $\sigma(\mathbf{A})$, satisfying $\lambda_i > \lambda_{i+1},\ i = 1, 2, \cdots, N-1$. Then

$$\lambda_{N-k} \leq \langle \mathbf{Ax}, \mathbf{x}\rangle \leq \lambda_1,\ \mathbf{x} \in S_{N-k},\ \|\mathbf{x}\|_2 = 1 \qquad (1)$$

with equality holding on the left hand side of (1) when $\mathbf{x} \in N(\mathbf{A} - \lambda_{N-k}\mathbf{I})$ and on the right hand side when $\mathbf{x} \in N(\mathbf{A} - \lambda_1 \mathbf{I})$. Also

$$\lambda_N \leq \langle \mathbf{Ax},\, \mathbf{x}\rangle \leq \lambda_{k+1},\ \mathbf{x} \in \hat{S}_{k+1},\ \|\mathbf{x}\|_2 = 1 \qquad (2)$$

with equality holding on the left hand side of (2) when $\mathbf{x} \in N(\mathbf{A} - \lambda_N \mathbf{I})$ and on the right hand side when $\mathbf{x} \in N(\mathbf{A} - \lambda_{k+1}\mathbf{I})$.

Proof: We shall only provide a prove for inequality (1), since the proof for inequality (2) follows in a similar manner. The eigendecomposition of $\mathbf{A}$ as given by the spectral theorem [7] is

$$\mathbf{A} = \sum_{i=1}^{N} \lambda_i \mathbf{G}_i \qquad (3)$$

where

$$\mathbf{G}_i = \sum_{j=1}^{m_i} \mathbf{u}_i^j (\mathbf{u}_i^j)^t \qquad (4)$$

are the orthogonal projectors onto $N(\mathbf{A} - \lambda_i \mathbf{I})$ along the range $R(\mathbf{A} - \lambda_i \mathbf{I})$, with the property that $\mathbf{G}_i \mathbf{G_j} = \delta_{ij} \mathbf{G}_i$,

where δ_{ij}, denotes the Kronecker delta. Furthermore, $\mathbf{I} = \sum_{i=1}^{N} \mathbf{G}_i$. Now, for $\mathbf{x} \in S_{N-k}$ and $\|\mathbf{x}\|_2 = 1$ implies that

$$\mathbf{x} = \sum_{i=1}^{N-k} \mathbf{G}_i \mathbf{x}.$$

Taking the inner product $\langle \mathbf{x}, \mathbf{x} \rangle$, we get

$$\begin{aligned} \langle \mathbf{x}, \mathbf{x} \rangle &= \sum_{i=1}^{N-k} \langle \mathbf{G}_i \mathbf{x}, \mathbf{x} \rangle \\ &= 1. \end{aligned}$$

Hence, the inner product $\langle \mathbf{A}\mathbf{x}, \mathbf{x} \rangle$, yields that

$$\begin{aligned} \langle \mathbf{A}\mathbf{x}, \mathbf{x} \rangle &= \left\langle \sum_{i=1}^{N-k} \lambda_i \mathbf{G}_i \mathbf{x}, \sum_{j=1}^{N-k} \mathbf{G}_j \mathbf{x} \right\rangle \\ &= \sum_{i=1}^{N-k} \sum_{j=1}^{N-k} \lambda_i \langle \mathbf{G}_i \mathbf{x}, \mathbf{G}_j \mathbf{x} \rangle \\ &= \sum_{i=1}^{N-k} \lambda_i \langle \mathbf{G}_i \mathbf{x}, \mathbf{x} \rangle. \end{aligned} \tag{5}$$

It follows from (5) that

$$\langle \mathbf{A}\mathbf{x}, \mathbf{x} \rangle \leq \lambda_1 \sum_{i=1}^{N-k} \langle \mathbf{G}_i \mathbf{x}, \mathbf{x} \rangle = \lambda_1$$

and

$$\langle \mathbf{A}\mathbf{x}, \mathbf{x} \rangle \geq \lambda_{N-k} \sum_{i=1}^{N-k} \langle \mathbf{G}_i \mathbf{x}, \mathbf{x} \rangle = \lambda_{N-k}.$$

If $\mathbf{x} \in N(\mathbf{A} - \lambda_{N-k}\mathbf{I})$, then

$$\mathbf{x} = \mathbf{G}_{N-k}\mathbf{x}$$

and the inner product $\langle \mathbf{A}\mathbf{x}, \mathbf{x} \rangle$,

$$\begin{aligned} \langle \mathbf{A}\mathbf{x}, \mathbf{x} \rangle &= \langle \lambda_{N-k} \mathbf{G}_{N-k}\mathbf{x}, \mathbf{x} \rangle \\ &= \lambda_{N-k} \langle \mathbf{x}, \mathbf{x} \rangle \\ &= \lambda_{N-k} \end{aligned}$$

If $\mathbf{x} \in N(\mathbf{A} - \lambda_1\mathbf{I})$, then

$$\mathbf{x} = \mathbf{G}_1\mathbf{x}$$

and the inner product $\langle \mathbf{A}\mathbf{x}, \mathbf{x} \rangle$,

$$\begin{aligned} \langle \mathbf{A}\mathbf{x}, \mathbf{x} \rangle &= \langle \lambda_1 \mathbf{G}_1\mathbf{x}, \mathbf{x} \rangle \\ &= \lambda_1 \langle \mathbf{x}, \mathbf{x} \rangle \\ &= \lambda_1. \end{aligned}$$

■

We note that, when $k = 0$ in (1) and (2) that,

$$\lambda_N \leq \langle \mathbf{A}\mathbf{x}, \mathbf{x} \rangle \leq \lambda_1, \ \|\mathbf{x}\|_2 = 1, \ S_N = \hat{S}_1 = \mathbb{R}^n.$$

Furthermore, we observe that span$\{S\}$, S_j and $\hat{S}_j$ are $\mathbf{A}$ and $\mathbf{A} - \lambda\mathbf{I}$ ($\lambda \in \mathbb{R}$) invariant subspaces.

Let $\mathbf{u}$ be a normalized eigenvector of $\mathbf{A}$ with corresponding eigenvalue λ, and let $\hat{\mathbf{u}}$ be a normalized perturbation of order ε of $\mathbf{u}$. Therefore $\hat{\mathbf{u}} = \frac{\mathbf{u}+\varepsilon\mathbf{u}^{\perp}}{\sqrt{1+\varepsilon^2}}$, where $\langle \mathbf{u}^{\perp}, \mathbf{u} \rangle = 0$ and $\|\mathbf{u}^{\perp}\|_2 = \|\hat{\mathbf{u}}\|_2 = 1$ (Note that it only suffices to consider perturbations perpendicular to $\mathbf{u}$).

Theorem 2: If $\mathbf{u}$ is a normalized eigenvector of $\mathbf{A}$ with corresponding eigenvalue λ, and if $\hat{\mathbf{u}}$ is the normalized perturbation of order ε of $\mathbf{u}$. Then the corresponding perturbation in λ is $\mathcal{O}(\varepsilon^2)$.

Proof: Taking the inner product of $\mathbf{A}\hat{\mathbf{u}}$ with $\hat{\mathbf{u}}$, we get

$$\begin{aligned} \hat{\lambda} &= \langle \mathbf{A}\hat{\mathbf{u}}, \hat{\mathbf{u}} \rangle \\ &= \frac{\langle \lambda\mathbf{u} + \varepsilon\mathbf{A}\mathbf{u}^{\perp}, \mathbf{u} + \varepsilon\mathbf{u}^{\perp} \rangle}{1+\varepsilon^2} \\ &= \frac{\lambda + 2\varepsilon\lambda\langle \mathbf{u}^{\perp}, \mathbf{u} \rangle + \varepsilon^2 \langle \mathbf{A}\mathbf{u}^{\perp}, \mathbf{u}^{\perp} \rangle}{1+\varepsilon^2} \\ &= \frac{\lambda(1+\varepsilon^2) + (\langle \mathbf{A}\mathbf{u}^{\perp}, \mathbf{u}^{\perp} \rangle - \lambda)\varepsilon^2}{1+\varepsilon^2} \\ &= \lambda + \mathcal{O}(\varepsilon^2). \end{aligned} \tag{6}$$

(7)

■

Theorem 3: Let $\mathbf{A}$ be a symmetric matrix with dominant eigenvalue λ_1 and $\mathbf{G}_1$ be the corresponding projector onto $N(\mathbf{A} - \lambda_1\mathbf{I})$. Then $\frac{\mathbf{A}^k}{\lambda_1^k}$ converges linearly to $\mathbf{G}_1$ with asymptotic error constant given by $\left|\frac{\lambda_2}{\lambda_1}\right|$.

Proof: For the k_{th} power of matrix $\mathbf{A}$, we get

$$\begin{aligned} \mathbf{A}^k &= \sum_{i=1}^{N} \lambda_i^k \mathbf{G}_i \\ &= \lambda_1^k \mathbf{G}_1 + \sum_{i=2}^{N} \lambda_i^k \mathbf{G}_i. \end{aligned}$$

Let

$$\begin{aligned} \mathbf{e}_k &= \frac{\mathbf{A}^k}{\lambda_1^k} - \mathbf{G}_1 \\ &= \sum_{i=2}^{N} \left(\frac{\lambda_i}{\lambda_1}\right)^k \mathbf{G}_i, \end{aligned} \tag{8}$$

then (8) is a spectral decomposition of the matrix $\mathbf{e}_k$. Hence, the spectrum is given by

$$\sigma(\mathbf{e}_k) = \left\{ \left(\frac{\lambda_2}{\lambda_1}\right)^k, \left(\frac{\lambda_3}{\lambda_1}\right)^k, \cdots, \left(\frac{\lambda_N}{\lambda_1}\right)^k, 0 \right\}$$

and $\|\mathbf{e}_k\|_2 = \left|\frac{\lambda_2}{\lambda_1}\right|^k < 1$. It follows that, $\lim_{k\to\infty} \|\mathbf{e}_k\|_2 = 0$ and $\mathbf{e}_k \to \mathbf{0}$. Furthermore, we get $\frac{\|\mathbf{e}_{k+1}\|_2}{\|\mathbf{e}_k\|_2} = \left|\frac{\lambda_2}{\lambda_1}\right|$. So convergence is linear with asymptotic error constant $\left|\frac{\lambda_2}{\lambda_1}\right|$. The latter is sometimes referred to as the convergence rate.

■

If $\mathbf{A}$ is replaced by $\mathbf{A}^m$ in Theorem 3 then the convergence is linear at the rate $\left|\frac{\lambda_2}{\lambda_1}\right|^m$.

III. The power method

Consider the iterative process

$$\mathbf{x}_k = \frac{\mathbf{A}\mathbf{x}_{k-1}}{\|\mathbf{A}\mathbf{x}_{k-1}\|_2}, \ \mathbf{x}_0 \in \mathbb{R}^n \tag{9}$$

where $\|\mathbf{x}_0\|_2 = 1$ and $\mathbf{G}_1\mathbf{x}_0 \neq \mathbf{0}$.

Then taking the limit as $k \to \infty$, we get

$$\mathbf{x}_k = \frac{\mathbf{A}^k\mathbf{x}_0}{\|\mathbf{A}^k\mathbf{x}_0\|_2} \tag{10}$$

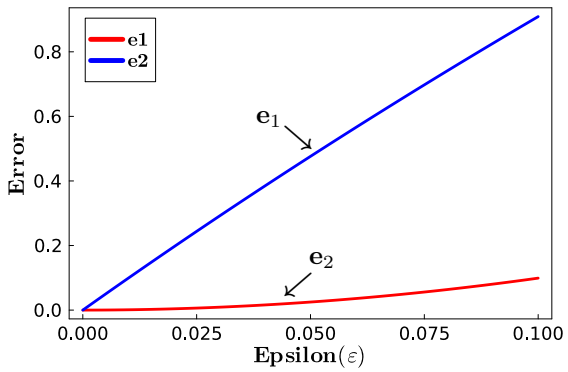

Fig. 1. Error vs ε for Example 1

and by Theorem 3

$$\mathbf{x}_k \to \frac{\lambda_1^k \mathbf{G}_1 \mathbf{x}_0}{|\lambda_1|^k \|\mathbf{G}_1 \mathbf{x}_0\|_2} = \mathbf{u}_1 \in N(\mathbf{A} - \lambda_1 \mathbf{I}).$$

It follows that in the limit as $k \to \infty$,

$$\mathbf{A}\mathbf{x}_k \to \lambda_1 \mathbf{u}_1$$

and

$$\frac{(\mathbf{A}\mathbf{x}_k)_p}{(\mathbf{x}_k)_p} \to \lambda_1$$

where $(\mathbf{x}_k)_p$ denotes the p_{th} component of $\mathbf{x}_k$ such that $\|\mathbf{x}_k\|_\infty = |(\mathbf{x}_k)_p|$. For k large enough, we write

$$\hat{\mathbf{u}}_1 = \mathbf{x}_k = \frac{\mathbf{u}_1 + \varepsilon \mathbf{u}_1^\perp}{\sqrt{1+\varepsilon^2}},$$

where $\varepsilon \to 0$ as $k \to \infty$ so that

$$\begin{aligned} \hat{\lambda}_1 &= \frac{(\mathbf{A}(\mathbf{u}_1 + \varepsilon \mathbf{u}_1^\perp))_p}{(\mathbf{u}_1 + \varepsilon \mathbf{u}_1^\perp)_p} \qquad (11) \\ &= \frac{(\lambda_1 \mathbf{u}_1 + \varepsilon \mathbf{A}\mathbf{u}_1^\perp)_p}{(\mathbf{u}_1 + \varepsilon \mathbf{u}_1^\perp)_p} \\ &= \frac{\lambda_1 + \dfrac{\varepsilon (\mathbf{A}\mathbf{u}_1^\perp)_p}{(\mathbf{u}_1)_p}}{1 + \dfrac{\varepsilon (\mathbf{u}_1^\perp)_p}{(\mathbf{u}_1)_p}} \\ &= \lambda_1 + \mathcal{O}(\varepsilon). \qquad (12) \end{aligned}$$

Since $\langle \mathbf{A}\mathbf{x}_k, \mathbf{x}_k \rangle = \lambda_1 + \mathcal{O}(\varepsilon^2)$ by theorem 2, it follows that the Rayleigh number $\langle \mathbf{A}\mathbf{x}_k, \mathbf{x}_k \rangle$ is a better approximation to λ_1 than (12).

Example 1: Consider the matrix

$$\mathbf{A} = \begin{bmatrix} 8 & 4 & 4 & 1 \\ 4 & 8 & 1 & 4 \\ 4 & 1 & 8 & 4 \\ 1 & 4 & 4 & 8 \end{bmatrix}$$

where $\lambda_1 = 17$ and $\mathbf{u} = \mathbf{u}_1 = [\frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}]^t$. We choose $\mathbf{u}^\perp = \mathbf{u}_1^\perp = [-\frac{1}{2}, \frac{1}{2}, -\frac{1}{2}, \frac{1}{2}]^t$ and plot the errors e1 $= |\lambda - \hat{\lambda}|$ obtained from equations (6), e2 $= |\lambda - \hat{\lambda}_1|$ obtained from equations (11), versus ε in Figure 1. The linear and quadratic dependence on ε is clearly seen from the blue and red plots, agreeing with equations (12) and (7), respectively.

Algorithm 1: **power**$(\mathbf{A})$

1: choose $\mathbf{x}_0 \in \mathbb{R}^n$ randomly, $\|\mathbf{x}_0\|_2 = 1$.
2: **for** $k = 1$ to K **do**
3: $\quad \mathbf{x}_1 = \mathbf{A}\mathbf{x}_0$
4: $\quad \mathbf{x}_1 = \frac{\mathbf{x}_1}{\|\mathbf{x}_1\|_2}$
5: $\quad$ **if** $\|\mathbf{x}_1 - \mathbf{x}_0\|_2 > \varepsilon$ **then**
6: $\quad\quad \mathbf{x}_0 = \mathbf{x}_1$
7: $\quad$ **else**
8: $\quad\quad \lambda_1 = \langle \mathbf{A}\mathbf{x}_1, \mathbf{x}_1 \rangle$
9: $\quad\quad \mathbf{u}_1 = \mathbf{x}_1$
10: $\quad\quad$ **stop**
11: $\quad$ **end if**
12: **end for**
13: **return** $\mathbf{u} = \mathbf{u}_1, \lambda = \lambda_1$

It is necessary to scale the iterates in step 4 to prevent numerical overflow.

Shifted iteration

Given $p \in \mathbb{R}$ one may determine the eigenvalue of $\mathbf{A}$ furthest away from p by applying the power method to $\mathbf{A} - p\mathbf{I}$. This is particularly useful for positive definite symmetric matrices, when the average of the eigenvalues is skewed towards λ_1. This allows the determination of the smallest mode $(\lambda_N, \mathbf{u}_N)$ by avoiding inverse iteration. In this case it is indeed true that

$$\tfrac{\text{trace}(\mathbf{A})}{n} - \lambda_N > \lambda_1 - \tfrac{\text{trace}(\mathbf{A})}{n},$$

where the latter two expressions are the first and second dominant eigenvalues of

$$\mathbf{A} - \tfrac{\text{trace}(\mathbf{A})}{n}\mathbf{I}.$$

Thus the power method on the latter matrix will yield the last mode. Let us find the optimal value of p, such that $\lambda_i - p > \lambda_{i+1} - p, i = 1, 2, \ldots, N-1$, with the requirement that the asymptotic error constant satisfies

$$\frac{\lambda_2 - p}{\lambda_1 - p} < \frac{\lambda_2}{\lambda_1}. \qquad (13)$$

We have the following cases that restrict the values of p:

(a) $p \in (-\infty, 0)$ contradicts (13),
(b) $p \in (\lambda_2, \infty)$, then $\lambda_N - p$ is dominant,
(c) $p \in (\frac{\lambda_N + \lambda_2}{2}, \lambda_2)$, then $\lambda_2 - p$ is no longer the second dominant eigenvalue.

Thus, for any $p \in (0, \frac{\lambda_N + \lambda_2}{2}]$, results in a smaller asymptotic error constant. However, $f(p) = \frac{\lambda_2 - p}{\lambda_1 - p}$ decreases on this interval, thus the optimal value is given by $p = \frac{\lambda_N + \lambda_2}{2}$. It is obvious that to use this result, reasonable approximations of λ_N and λ_2 are required. In other words, with this value of p, the power method using $\mathbf{A} - p\mathbf{I}$ will converge faster than the power method using $\mathbf{A}$, to the first mode.

Example 2: Consider the matrix

$$\mathbf{A} = \begin{bmatrix} 7 & 4 & 3 & 2 & 1 \\ 4 & 8 & 0 & 4 & 3 \\ 3 & 0 & 9 & 6 & 5 \\ 2 & 4 & 6 & 10 & 7 \\ 1 & 3 & 5 & 7 & 11 \end{bmatrix},$$

where $\lambda_1 = 24.406875$ to six decimal digits. Figure III illustrates the convergence for values of $p = 0, 5.2, 9$. The red curve clearly indicates that the optimal value of p for convergence is $p = 5.2$, which is obtained from the exact eigenvalues λ_2 and λ_N. The values of $p = 9$ is obtained from $\frac{\text{trace}(\mathbf{A})}{N}$.

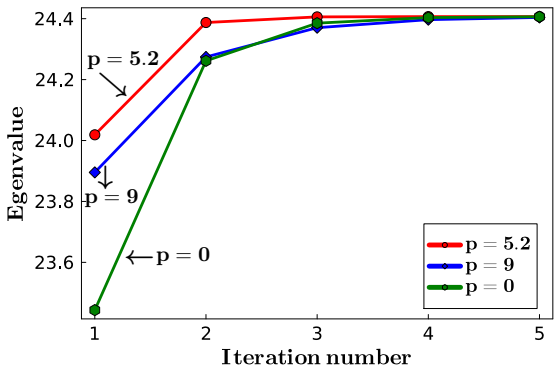

Fig. 2. Error vs ε for Example 2

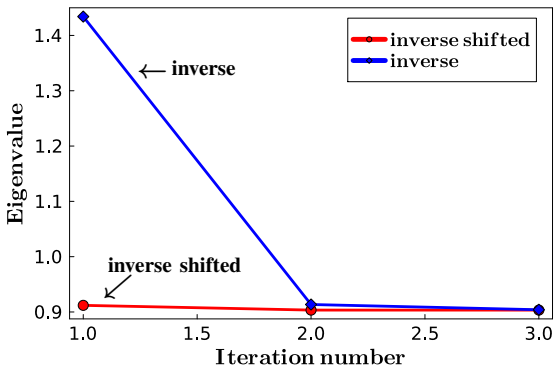

Fig. 3. Eigenvalue vs iteration number for Example 3

IV. Inverse iteration

In this section, we assume that λ_{N-1} is dominant to λ_N. It is clear that the power method applied to $\mathbf{A}^{-1}$ will yield the dominant eigenvalue of $\mathbf{A}^{-1}$, equivalently the reciprocal of the least dominant eigenvalue of $\mathbf{A}$. Thus we are able to recover the N_{th} eigenmode and the corresponding rate of convergence is $\left|\frac{\lambda_N}{\lambda_{N-1}}\right|$.

Algorithm 2: **Inverse iteration**

1: choose $\mathbf{x}_0 \in \mathbb{R}^n$ randomly, $\|\mathbf{x}_0\|_2 = 1$.
2: **for** $k = 1$ to K **do**
3: solve $\mathbf{x}_0 = \mathbf{A}\mathbf{x}_1$ for $\mathbf{x_1}$
4: $\mathbf{x}_1 = \frac{\mathbf{x}_1}{\|\mathbf{x}_1\|_2}$
5: **if** $\|\mathbf{x}_1 - \mathbf{x}_0\|_2 > \varepsilon$ **then**
6: $\mathbf{x}_0 = \mathbf{x}_1$
7: **else**
8: $\lambda_N = \langle \mathbf{A}\mathbf{x}_1, \mathbf{x}_1 \rangle$
9: $\mathbf{u}_N = \mathbf{x}_1$
10: stop
11: **end if**
12: **end for**

Inverse power iteration is more costly than the power method due to the solution of a linear system at step 3, however, a LU decomposition can minimize this cost.

Inverse shifted iteration

If an approximation $\hat{\lambda}_k$ to an eigenvalue λ_k is known then the inverse power method on $\mathbf{A} - \hat{\lambda}_k\mathbf{I}$ yields the least dominant eigenvalue of $\mathbf{A} - \hat{\lambda}_k\mathbf{I}$, namely $\lambda_k - \hat{\lambda}_k$. The closer $\hat{\lambda}_k$ is to λ_k, the faster the convergence. Such approximations to λ_k (especially λ_1, λ_N) are found in the literature [10]. However, if the eigenvalues are clustered near λ_k then it is likely to converge to the wrong eigenmode, unless $\hat{\lambda}_k$ is a very good approximation to λ_k.

Example 3: For the matrix of Example 2, $\lambda_N = 0.903405$ correct to six decimal places. Fig. 3 shows that inverse shifted iteration on $\mathbf{A} - \mathbf{I}$ converges after one iteration, compared to a plain inverse iteration.

V. Deflation

We may determine all the eigenmodes of $\mathbf{A}$ by the process of deflation, by having obtained the first mode $(\lambda_1, \mathbf{u}_1^1)$ by the power method, we will then determine all the other modes corresponding to λ_1. It follows from (3) and (4), that

$$\mathbf{A} - \lambda_1 \mathbf{u}_1^1 (\mathbf{u}_1^1)^t = \sum_{j=2}^{m_1} \lambda_1 \mathbf{u}_1^j (\mathbf{u}_1^j)^t + \sum_{i=2}^{N} \lambda_i \mathbf{G}_i. \tag{14}$$

Hence, the spectrum of $(\mathbf{A} - \lambda_1 \mathbf{u}_1^1 (\mathbf{u}_1^1)^t)$ is given by

$$\sigma(\mathbf{A} - \lambda_1 \mathbf{u}_1^1 (\mathbf{u}_1^1)^t) = \{0,\, \lambda_1,\, \lambda_2,\, \cdots,\, \lambda_N\},$$

where λ_1 and zero are eigenvalues of algebraic multiplicity $m_1 - 1$ and unity respectively. The power method applied to $\mathbf{B}_1 = \mathbf{A} - \lambda_1 \mathbf{u}_1^1 (\mathbf{u}_1^1)^t$ will yield an eigenvalue $\mathbf{u}$ corresponding to λ_1. From (14) we note that

$$(\mathbf{A} - \lambda_1 \mathbf{u}_1^1 (\mathbf{u}_1^1)^t)\mathbf{x}_0 = \sum_{j=2}^{m_1} \lambda_1 \langle \mathbf{x}_0, \mathbf{u}_1^j \rangle \mathbf{u}_1^j + \sum_{i=2}^{N} \lambda_i \mathbf{G}_i \mathbf{x}_0.$$

Hence, $\mathbf{u} \in \text{span}\{\mathbf{u}_1^j\}_{j=2}^{m_1}$ implies that $\mathbf{u} \perp \mathbf{u}_1$. Let $\mathbf{u}_1^2 = \mathbf{u}$ and proceed with further deflation, having obtained $\{\mathbf{u}_1^j\}_{j=1}^{k-1}$, $k = 2, 3, \cdots, m_1$, $\mathbf{u}_1^k$ is obtained by applying the power method to

$$\begin{aligned}\mathbf{B}_{k-1} &= \mathbf{A} - \lambda_1 \sum_{j=1}^{k-1} \mathbf{u}_1^j (\mathbf{u}_1^j)^t \\ &= \lambda_1 \sum_{j=k}^{m_1} \mathbf{u}_1^j (\mathbf{u}_1^j)^t + \sum_{i=2}^{N} \lambda_i \mathbf{G}_i.\end{aligned}$$

Clearly, the spectrum of $\mathbf{B}_{k-1}$ is given by $\sigma(\mathbf{B}_{k-1}) = \{0, \lambda_1, \lambda_2, \cdots, \lambda_N\}$ with λ_1 having algebraic multiplicity $m_1 - k + 1$ and zero having algebraic multiplicity $k - 1$. Deflation of λ_1 is terminated with the power method on $\mathbf{B}_{m_1-1}$ yielding $\mathbf{u}_1^{m_1}$, thus $\mathbf{G}_1$ is obtained. Having obtained $(\lambda_1, \mathbf{G}_1)$, we have from (3)

$$\begin{aligned}\mathbf{B}_{m_1} &= \mathbf{A} - \lambda_1 \mathbf{G}_1 \\ &= \sum_{i=2}^{N} \lambda_i \mathbf{G}_i\end{aligned}$$

and

$$\sigma(\mathbf{A} - \lambda_1 \mathbf{G}_1) = \{0,\, \lambda_2,\, \lambda_3,\, \lambda_N\}.$$

The power method applied to $\mathbf{B}_{m_1}$, will converge to $(\lambda_2, \mathbf{u}_2^1)$. Repetition of the deflation process described above will

result in the complete deflation of λ_2. Thus all modes up to $(\lambda_N, \mathbf{u}_N^{m_N})$ can be determined. Clearly,

$$\begin{aligned}\mathbf{B}_{m_k} &= \mathbf{A} - \sum_{i=1}^{k} \lambda_i \mathbf{G}_i \\ &= \sum_{i=k+1}^{N} \lambda_i \mathbf{G}_i.\end{aligned}$$

Thus $\mathbf{x}_0 \in \hat{S}_{k+1}$ which implies that $\mathbf{B}_{m_k}\mathbf{x}_0 \in \hat{S}_{k+1}$ and $\hat{S}_{k+1}$ is $\mathbf{B}_{m_k}$ invariant.

Algorithm 3: **Deflation**

call **power**($\mathbf{A}$)
$m = 1$
for $i = 1$ to $n-1$ **do**
  $\mathbf{z} = \mathbf{u}$
  $\beta = \lambda$
  output $m,\ \beta,\ \mathbf{z}$
  $\mathbf{A} = \mathbf{A} - \beta \mathbf{z}\mathbf{z}^t$
  call power($\mathbf{A}$)
  if $|\beta - \lambda| \leq \varepsilon$ **then**
    $m = m + 1$
  else
    $m = 1$
  end if
end for

Theorem 4: Let $(\hat{\lambda}, \hat{\mathbf{u}})$ approximate the eigenmode $(\lambda, \mathbf{u})$, where $\hat{\mathbf{u}}$ approximates $\mathbf{u}$ to order $\mathcal{O}(\varepsilon)$. Then

$$\begin{aligned}\hat{\mathbf{B}}^k\mathbf{x}_0 &= \mathbf{B}^k\mathbf{x}_0 \\ &\quad - \lambda\varepsilon\left(\langle \mathbf{x}_0,\, \mathbf{u}\rangle \mathbf{B}^{k-1}\mathbf{u}^\perp + \langle \mathbf{B}^{k-1}\mathbf{x}_0, \mathbf{u}^\perp\rangle \mathbf{u}\right), \quad (15)\end{aligned}$$

where $\hat{\mathbf{B}} = \mathbf{A} - \hat{\lambda}\hat{\mathbf{u}}\hat{\mathbf{u}}^t$, is the actual deflation matrix and $\mathbf{B} = \mathbf{A} - \lambda\mathbf{u}\mathbf{u}^t$ is the exact deflation matrix.

Proof: From theorem 2, we have that

$$\hat{\mathbf{u}} = \frac{\mathbf{u} + \varepsilon\mathbf{u}^\perp}{\sqrt{1+\varepsilon^2}}$$

and that $\hat{\lambda} = \lambda + \mathcal{O}(\varepsilon^2)$. Thus, we get

$$\begin{aligned}\hat{\lambda}\hat{\mathbf{u}}\hat{\mathbf{u}}^t &= \left(\lambda + \mathcal{O}(\varepsilon^2)\right) \frac{(\mathbf{u}+\varepsilon\mathbf{u}^\perp)(\mathbf{u}+\varepsilon\mathbf{u}^\perp)^t}{1+\varepsilon^2} \\ &= \lambda\mathbf{u}\mathbf{u}^t + \lambda\varepsilon\left(\mathbf{u}^\perp\mathbf{u}^t + \mathbf{u}(\mathbf{u}^\perp)^t\right) \quad \text{to order } \varepsilon.\end{aligned}$$

Hence, we obtain

$$\mathbf{A} - \hat{\lambda}\hat{\mathbf{u}}\hat{\mathbf{u}}^t = \mathbf{A} - \lambda\mathbf{u}\mathbf{u}^t - \lambda\varepsilon\left(\mathbf{u}^\perp\mathbf{u}^t + \mathbf{u}(\mathbf{u}^\perp)^t\right).$$

It follows that

$$\hat{\mathbf{B}}\mathbf{x}_0 = \mathbf{B}\mathbf{x}_0 - \lambda\varepsilon\left(\langle \mathbf{x}_0,\, \mathbf{u}\rangle \mathbf{u}^\perp + \langle \mathbf{x}_0, \mathbf{u}^\perp\rangle \mathbf{u}\right). \quad (16)$$

Hence, statement (15) is true for $k = 1$. Assume that (15) is true for k, then for $k+1$, we proceed as follows. Replace $\mathbf{x}_0$ by $\hat{\mathbf{B}}\mathbf{x}_0$ in (15) to get

$$\begin{aligned}\hat{\mathbf{B}}^{k+1}\mathbf{x}_0 &= \mathbf{B}^k\hat{\mathbf{B}}\mathbf{x}_0 \\ &\quad - \lambda\varepsilon\left(\langle \hat{\mathbf{B}}\mathbf{x}_0,\, \mathbf{u}\rangle \mathbf{B}^{k-1}\mathbf{u}^\perp + \langle \mathbf{B}^{k-1}\hat{\mathbf{B}}\mathbf{x}_0, \mathbf{u}^\perp\rangle \mathbf{u}\right). \quad (17)\end{aligned}$$

Note from (16) that

$$\begin{aligned}\mathbf{B}^k\hat{\mathbf{B}}\mathbf{x}_0 &= \mathbf{B}^{k+1}\mathbf{x}_0 - \lambda\varepsilon\left(\langle \mathbf{x}_0,\, \mathbf{u}\rangle \mathbf{B}^k\mathbf{u}^\perp + \langle \mathbf{x}_0,\, \mathbf{u}\rangle \mathbf{B}^k\mathbf{u}\right) \\ &= \mathbf{B}^{k+1}\mathbf{x}_0 - \lambda\varepsilon\langle \mathbf{x}_0,\, \mathbf{u}\rangle \mathbf{B}^k\mathbf{u}^\perp. \quad (18)\end{aligned}$$

Furthermore, from (16)

$$\begin{aligned}&\langle \hat{\mathbf{B}}\mathbf{x}_0, \mathbf{u}\rangle \\ &= \langle \mathbf{B}\mathbf{x}_0, \mathbf{u}\rangle - \lambda\varepsilon\left(\langle \mathbf{x}_0,\, \mathbf{u}\rangle\langle \mathbf{u}^\perp, \mathbf{u}\rangle + \langle \mathbf{x}_0, \mathbf{u}^\perp\rangle\langle \mathbf{u}, \mathbf{u}\rangle\right) \\ &= -\lambda\varepsilon\langle \mathbf{x}_0,\, \mathbf{u}^\perp\rangle. \quad (19)\end{aligned}$$

To arrive at (18) and (19) we have used the fact that $\mathbf{B}$ is symmetric and $\mathbf{u} \in N(\mathbf{B})$. From (16)

$$\mathbf{B}^{k-1}\hat{\mathbf{B}}\mathbf{x}_0 = \mathbf{B}^k\mathbf{x}_0 - \lambda\varepsilon\langle \mathbf{x}_0,\, \mathbf{u}\rangle \mathbf{B}^{k-1}\mathbf{u}^\perp$$

so that

$$\langle \mathbf{B}^{k-1}\hat{\mathbf{B}}\mathbf{x}_0, \mathbf{u}^\perp\rangle = \langle \mathbf{B}^k\mathbf{x}_0, \mathbf{u}^\perp\rangle - \lambda\varepsilon\langle \mathbf{x}_0,\, \mathbf{u}\rangle\langle \mathbf{B}^{k-1}\mathbf{u}^\perp, \mathbf{u}^\perp\rangle. \quad (20)$$

Using (18)-(20) in (17), we finally have

$$\begin{aligned}\hat{\mathbf{B}}^{k+1}\mathbf{x}_0 &= \mathbf{B}^{k+1}\mathbf{x}_0 \\ &\quad - \lambda\varepsilon\left(\langle \mathbf{x}_0,\, \mathbf{u}\rangle \mathbf{B}^k\mathbf{u}^\perp + \langle \mathbf{B}^k\mathbf{x}_0, \mathbf{u}^\perp\rangle \mathbf{u}\right) + \mathcal{O}(\varepsilon).\end{aligned}$$

■

Apart from normalization of the iterates, we observe from (15), that if $\mathbf{B}^k\mathbf{x}_0$ converges to an eigenvector $\mathbf{v}$ of $\mathbf{A}$, then $\hat{\mathbf{B}}^k\mathbf{x}_0$ converges to an eigenvector $\hat{\mathbf{v}}$ of $\mathbf{A}$. Clearly, $\mathbf{B}^{k-1}\mathbf{u}^\perp$ converges to say $\mathbf{b}$, where $\mathbf{b} \perp \mathbf{u}$. Thus we write

$$\begin{aligned}\hat{\mathbf{v}} &= \mathbf{v} - \lambda\varepsilon\langle \mathbf{x}_0, \mathbf{u}\rangle \mathbf{b} \\ &= \mathbf{v} + \mathcal{O}(\varepsilon)\end{aligned}$$

Thus $\hat{\mathbf{v}}$ approximates $\mathbf{v}$ to $\mathcal{O}(\varepsilon)$ and the deflation process is stable.

VI. Circumventing deflation

A. What doesn't work

Suppose that we have determined $(\hat{\lambda}_1, \hat{\mathbf{u}}_1^1)$ using the power method. Define
$S^\star = \text{span}\{(\hat{\mathbf{u}}_1^1)^\perp\} = \{\mathbf{x} \in \mathbb{R}^n | \langle \mathbf{x}, \hat{\mathbf{u}}_1^1\rangle = 0\}$. Then $\mathbf{x} \in S^\star$ which implies that

$$\begin{aligned}&\langle \mathbf{A}\mathbf{x}, \hat{\mathbf{u}}_1^1\rangle \\ &= \langle \mathbf{x}, \mathbf{A}\hat{\mathbf{u}}_1^1\rangle \\ &= \hat{\lambda}_1\langle \mathbf{x}, \hat{\mathbf{u}}_1^1\rangle \\ &= 0.\end{aligned}$$

Thus, $\mathbf{A}\mathbf{x} \in S^\star$ and $S^\star$ is $\mathbf{A}$ invariant. Therefore the power method applied to $\mathbf{x}_0^\star \in S^\star$ should converge to the next dominant mode $(\hat{\lambda}_1, \hat{\mathbf{u}}_1^2)$. However, this is not true in practice. We provide a brief analysis to understand why. It is nearly impossible to generate a $\mathbf{x}_0^\star \in S^\star$. Recall that $\hat{\lambda}_1 = \lambda_1 + \mathcal{O}(\varepsilon^2)$ and $\hat{\mathbf{u}}_1^1 = \frac{\mathbf{u}_1^1 + \varepsilon(\mathbf{u}_1^1)^\perp}{\sqrt{1+\varepsilon^2}}$. In the discussion that follows we shall work to $\mathcal{O}(\varepsilon)$. Given $\mathbf{x}_0 \in \mathbb{R}^n$ randomly chosen, we generate $\mathbf{x}_0^\star$ by Gram Schmidt orthogonalization. Hence, we obtain that

$$\begin{aligned}&\mathbf{x}_0^\star \\ &= \mathbf{x}_0 - \frac{\langle \mathbf{x}_0, \mathbf{u}_1^1 + \varepsilon(\mathbf{u}_1^1)^\perp\rangle(\mathbf{u}_1^1 + \varepsilon(\mathbf{u}_1^1)^\perp)}{1+\varepsilon^2} \\ &= \mathbf{x}_0 - \langle \mathbf{x}_0, \mathbf{u}_1^1\rangle\mathbf{u}_1^1 - \varepsilon\langle \mathbf{x}_0, (\mathbf{u}_1^1)^\perp\rangle\mathbf{u}_1^1 - \varepsilon\langle \mathbf{x}_0, \mathbf{u}_1^1\rangle(\mathbf{u}_1^1)^\perp. \quad (21)\end{aligned}$$

and

$$\mathbf{A}^k\mathbf{x}_0^\star$$

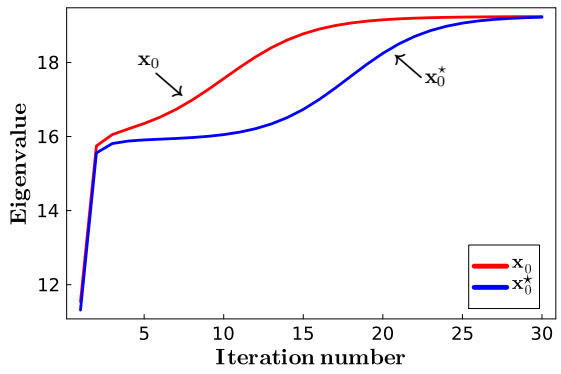

Fig. 4. Eigenvalue vs iteration number for Example 4

$$= \mathbf{A}^k\mathbf{x}_0 - \lambda_1^k\langle \mathbf{x}_0, \mathbf{u}_1^1\rangle \mathbf{u}_1^1 - \varepsilon\lambda_1^k\langle \mathbf{x}_0, (\mathbf{u}_1^1)^\perp\rangle \mathbf{u}_1^1 - \varepsilon\langle \mathbf{x}_0, \mathbf{u}_1^1\rangle \mathbf{A}^k(\mathbf{u}_1^1)^\perp. \tag{22}$$

From (14) and the spectral theorem, we have that

$$\mathbf{A}^k(\mathbf{u}_1^1)^\perp = \lambda_1^k \sum_{j=2}^{m_1}\langle \mathbf{u}_1^j, (\mathbf{u}_1^1)^\perp\rangle \mathbf{u}_1^j + \sum_{i=2}^{N}\lambda_i^k \mathbf{G}_i(\mathbf{u}_1^1)^\perp = \lambda_1^k \sum_{j=2}^{m_1}\langle \mathbf{u}_1^j, (\mathbf{u}_1^1)^\perp\rangle \mathbf{u}_1^j \quad \text{for } k \text{ large.} \tag{23}$$

Using (23), equation (22) becomes

$$\mathbf{A}^k\mathbf{x}_0^\star = \mathbf{A}^k\mathbf{x}_0 - \lambda_1^k\langle \mathbf{x}_0, \mathbf{u}_1^1\rangle \mathbf{u}_1^1 - \varepsilon\lambda_1^k\langle \mathbf{x}_0, (\mathbf{u}_1^1)^\perp\rangle \mathbf{u}_1^1 - \varepsilon\lambda_1^k\langle \mathbf{x}_0, \mathbf{u}_1^1\rangle \sum_{j=2}^{m_1}\langle \mathbf{u}_1^j, (\mathbf{u}_1^1)^\perp\rangle \mathbf{u}_1^j. \tag{24}$$

Since $\varepsilon\lambda_1^k << \lambda_1^k$, we may ignore the $\mathcal{O}(\varepsilon)$ terms in (24) to get

$$\begin{aligned}&\mathbf{A}^k\mathbf{x}_0^\star\\ &= \mathbf{A}^k\mathbf{x}_0 - \lambda_1^k\langle \mathbf{x}_0, \mathbf{u}_1^1\rangle \mathbf{u}_1^1\\ &= \mathbf{A}^k\mathbf{x}_0 - \lambda_1^k\langle \mathbf{x}_0, \mathbf{u}_1^1\rangle \hat{\mathbf{u}}_1^1 + \lambda_1^k\varepsilon\langle \mathbf{x}_0, \mathbf{u}_1^1\rangle (\mathbf{u}_1^1)^\perp\\ &= \mathbf{A}^k\mathbf{x}_0 - \lambda_1^k\langle \mathbf{x}_0, \mathbf{u}_1^1\rangle \hat{\mathbf{u}}_1^1,\end{aligned} \tag{25}$$

where we have ignored $\varepsilon\lambda_1^k$ as before. Since $\mathbf{A}^k\mathbf{x}_0$ in (25) converges to a vector parallel to $\hat{\mathbf{u}}_1^1$, we converge again to the first mode.

Example 4: The matrix

$$\mathbf{A} = \begin{bmatrix} 10 & 1 & 2 & 3 & 4\\ 1 & 9 & -1 & 2 & -3\\ 2 & -1 & 7 & 3 & -5\\ 3 & 2 & 3 & 12 & -1\\ 4 & -3 & -5 & -1 & 15 \end{bmatrix},$$

has eigenvalues $\lambda_1 = 19.242065$ and $\lambda_2 = 15.915101$, both of multiplicity one. The power method is applied with $\mathbf{x}_0$ and $\mathbf{x}_0^\star \perp \mathbf{x}_0$. In Fig. 4 it is clearly seen that the power method with $\mathbf{x}_0^\star$ converges to the first mode, rather than to the second mode.

B. What works

Suppose we have determined the modes $\{(\hat{\lambda}_1, \hat{\mathbf{u}}_1^j)\}_{j=1}^{p-1}$, $p \le m_1$. Now consider the iterative process

$$\mathbf{x}_0^\star = \mathbf{x}_0 - \sum_{j=1}^{p-1}\langle \mathbf{x}_0, \hat{\mathbf{u}}_1^j\rangle \hat{\mathbf{u}}_1^j$$

$$\mathbf{x}_k^\star = \mathbf{A}\mathbf{x}_{k-1}^\star - \hat{\lambda}_1 \sum_{j=1}^{p-1}\langle \mathbf{x}_{k-1}^\star, \hat{\mathbf{u}}_1^j\rangle \hat{\mathbf{u}}_1^j. \tag{26}$$

It follows from (26) that

$$\mathbf{x}_k^\star = \mathbf{A}^k\mathbf{x}_0 - \hat{\lambda}_1^k \sum_{j=1}^{p-1}\langle \mathbf{x}_0, \hat{\mathbf{u}}_1^j\rangle \hat{\mathbf{u}}_1^j$$

and the from the inner product of $\mathbf{x}_k^\star$ and $\hat{\mathbf{u}}_1^j$, yields that

$$\langle \mathbf{x}_k^\star, \hat{\mathbf{u}}_1^j\rangle = \langle \mathbf{x}_0, \mathbf{A}^k\hat{\mathbf{u}}_1^j\rangle - \hat{\lambda}_1^k\langle \mathbf{x}_0, \mathbf{u}_1^j\rangle = 0.$$

Hence, the iterates $\mathbf{x}_k^\star \in \text{span}\left\{\{(\hat{\mathbf{u}}_1^j)^\perp\}_{j=1}^{p-1}\right\}$, and so the **power method with orthogonalization** is forced to converge to the next mode $(\hat{\lambda}_1, \hat{\mathbf{u}}_1^p)$. Similarly having obtained

$$\{\mathbf{u}_i^j,\ \mathbf{u}_q^r | i = 1,\ 2, \cdots,\ q-1;\\ j = 1,\ 2, \cdots,\ m_i\,; r = 1, 2, \cdots, p-1\},$$

the iterative process

$$\mathbf{x}_k^\star = \mathbf{A}\mathbf{x}_{k-1}^\star - \sum_{i=1}^{q-1}\hat{\lambda}_i\sum_{j=1}^{m_i}\langle \mathbf{x}_{k-1}^\star, \hat{\mathbf{u}}_i^j\rangle \hat{\mathbf{u}}_i^j - \hat{\lambda}_q\sum_{r=1}^{p-1}\langle \mathbf{x}_{k-1}^\star, \hat{\mathbf{u}}_q^r\rangle \hat{\mathbf{u}}_q^r$$

will converge to the mode $\hat{\mathbf{u}}_q^p$. A simple algorithm to determine the second mode is presented in Algorithm 4

Algorithm 4: **orthogonalization**

call **power**($\mathbf{A}$) with output $\lambda_1, \mathbf{u}_1$
choose $\mathbf{x}_0 \in \mathbb{R}^n$ randomly, $\|\mathbf{x}_0\|_2 = 1$.
$\mathbf{x}_0 = \mathbf{x}_0 - \langle \mathbf{x}_0, \mathbf{u}_1\rangle \mathbf{u}_1$
for $k = 1$ to K **do**
  $\mathbf{x}_1 = \mathbf{A}\mathbf{x}_0 - \lambda_1\langle \mathbf{x}_0, \mathbf{u}_1\rangle \mathbf{u}_1$
  $\mathbf{x}_1 = \frac{\mathbf{x}_1}{\|\mathbf{x}_1\|_2}$
  if $\|\mathbf{x}_1 - \mathbf{x}_0\|_2 > \varepsilon$ **then**
    $\mathbf{x}_0 = \mathbf{x}_1$
  else
    $\lambda_2 = \langle \mathbf{A}\mathbf{x}_1, \mathbf{x}_1\rangle$
    $\mathbf{u}_2 = \mathbf{x}_1$
    stop
  end if
end for

Example 5: The power method with orthogonalization is applied to the matrix of Example 4, to determine the second mode. Results are depicted in Fig. 5. Compared to Fig. 4, we notice that we now have convergence to the second mode.

C. Projectors

We now consider a special case where projectors are useful. Consider the distribution

$$|\lambda_1| >> |\lambda_2| > |\lambda_3| > \cdots > |\lambda_{N-1}| > |\lambda_N|.$$

of the eigenvalues of $\mathbf{A}$. From Lagrange interpolation, we have that

$$\mathbf{G}_1 = \prod_{i=2}^{N}\frac{\mathbf{A} - \lambda_i\mathbf{I}}{\lambda_1 - \lambda_i},$$

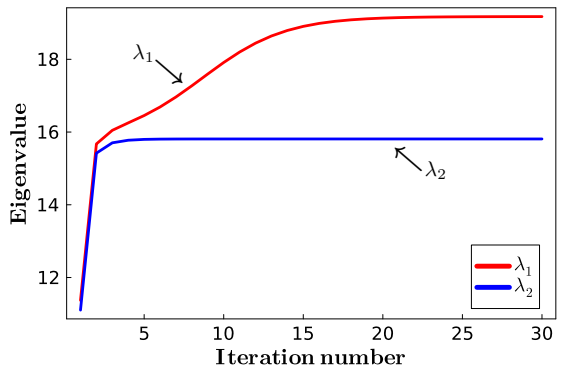

Fig. 5. Power method with orthogonalization for Example 5

so that $\mathbf{G}_1$ is approximately proportional to $\mathbf{A}^{N-1}$. Thus an eigenvector $\mathbf{u}_1^1$ is given by

$$\begin{aligned}\mathbf{u}_1^1 &= \frac{\mathbf{G}_1\mathbf{x}_0}{\|\mathbf{G}_1\mathbf{x}_0\|}\\ &\approx \frac{\mathbf{A}^{N-1}\mathbf{x}_0}{\|\mathbf{A}^{N-1}\mathbf{x}_0\|}\\ &= \hat{\mathbf{u}}_1^1\end{aligned}$$

and the Rayleigh number $\hat{\lambda}_1 = \langle \mathbf{A}\hat{\mathbf{u}}_1^1, \hat{\mathbf{u}}_1^1\rangle$ is a good approximation to λ_1. This may further be refined by a single application of shifted inverse iteration to provide a fairly accurate approximation $(\hat{\lambda}_1, \hat{\mathbf{u}}_1^1)$. Thus having obtained the first mode we may proceed to obtain a mode corresponding to λ_2. Note that

$$\mathbf{G}_2 = \prod_{\substack{i=1\\ i\neq 2}}^{N} \frac{\mathbf{A}-\lambda_i\mathbf{I}}{\lambda_2-\lambda_i}.$$

Thus $\mathbf{G}_2$ is approximately proportional to $(\mathbf{A}-\lambda_1\mathbf{I})\mathbf{A}^{N-2}$ and

$$\begin{aligned}\mathbf{u}_2^1 &= \frac{\mathbf{G}_2\mathbf{x}_0}{\|\mathbf{G}_2\mathbf{x}_0\|}\\ &\approx \frac{(\mathbf{A}-\lambda_1\mathbf{I})\mathbf{A}^{N-2}\mathbf{x}_0}{\|(\mathbf{A}-\lambda_1\mathbf{I})\mathbf{A}^{N-2}\mathbf{x}_0\|}\\ &= \hat{\mathbf{u}}_2^1.\end{aligned} \tag{27}$$

As before the Rayleigh number $\hat{\lambda}_2 = \langle \mathbf{A}\hat{\mathbf{u}}_2^1, \hat{\mathbf{u}}_2^1\rangle$ is an approximation to λ_2. This may be refined by a single or double step shifted inverse iteration to yield a fairly accurate approximation to the second mode. This is particularly suited to the case $m_i = 1$, namely $N = n$.

Example 6: Consider the matrix [5]

$$\begin{bmatrix} 5 & 7 & 6 & 5\\ 7 & 10 & 8 & 7\\ 6 & 8 & 10 & 9\\ 5 & 7 & 9 & 10\end{bmatrix} \tag{28}$$

where $\sigma(\mathbf{A}) = \{0.010150,\ 0.843907,\ 3.858057,\ 30.288685\}$ accurate to six decimal digits. The power method with $N-1 = 3$ iterations results in an error $|\hat{\lambda}_1 - \lambda_1| = 2.82\times 10^{-5}$. With one step of shifted iteration we achieve $|\hat{\lambda}_1-\lambda_1| = 5.45\times 10^{-12}$. Likewise for the second mode we obtain $|\hat{\lambda}_2-\lambda_2| = 6.44\times 10^{-5}$ using (27), $|\hat{\lambda}_2-\lambda_2| = 7.07\times 10^{-11}$ after one shifted inverse iteration and $|\hat{\lambda}_2-\lambda_2| = 7.11\times 10^{-15}$ after two shifted inverse iterations.

D. Distinct Eigenvalues

The following method is useful when $m_i = 1$, that is, the eigenvalues are distinct. Use the power method on $\mathbf{A}$ with $\mathbf{x}_0 \in \mathbb{R}^n$, chosen randomly and $\|\mathbf{x}_0\|_2 = 1$, to determine the first mode $(\lambda_{k_1}, \mathbf{u}_{k_1})$, $k_1 = 1$. Let $\{k_i\}_{i=2}^n \in \{2,\ 3,\ \cdots,\ n\}$, then apply the power method on $\mathbf{A}-\lambda_{k_1}\mathbf{I}$ to determine the next mode $(\lambda_{k_2}, \mathbf{u}_{k_2})$. Next apply the power method on $(\mathbf{A}-\lambda_{k_1}\mathbf{I})(\mathbf{A}-\lambda_{k_2}\mathbf{I})$ to determine the next mode $(\lambda_{k_3}, \mathbf{u}_{k_3})$. Having determined the modes $(\lambda_{k_i}, \mathbf{u}_{k_i})_{i=1}^{p-1}$, use the power method on

$$\tilde{\mathbf{B}}_p = \prod_{i=1}^{p-1}(\mathbf{A}-\lambda_{k_i}\mathbf{I}) \tag{29}$$

to determine $(\lambda_{k_p}, \mathbf{u}_{k_p})$. Note that $\tilde{\mathbf{B}}_p\mathbf{x}_0 \perp \{\mathbf{u}_{k_i}\}_{i=1}^{p-1}$ as

$$\begin{aligned}\langle \tilde{\mathbf{B}}\mathbf{x}_0,\ \mathbf{u}_{k_m}\rangle &= \langle \mathbf{x}_0,\ \prod_{i=1}^{p-1}(\mathbf{A}-\lambda_{k_i}\mathbf{I})\mathbf{u}_{k_m}\rangle \qquad 1\le m\le p-1\\ &= \langle \mathbf{x}_0,\ \prod_{\substack{i=1\\ i\neq m}}^{p-1}(\mathbf{A}-\lambda_{k_i}\mathbf{I})(\mathbf{A}\mathbf{u}_{k_m}-\lambda_{k_m}\mathbf{u}_{k_m})\rangle\\ &= 0.\end{aligned}$$

Thus all iterates are in a subspace perpendicular to $\text{span}\{\mathbf{u}_{k_i}\}_{i=1}^{p-1}$.

Lemma 1: Let $\tilde{\mathbf{B}}_p$ be defined as in (29), then

$$\tilde{\mathbf{B}}_p = \sum_{j=p}^{n}\prod_{i=1}^{p-1}(\lambda_{k_j}-\lambda_{k_i})\mathbf{G}_{k_j},\quad p>1.$$

Proof:

$$\begin{aligned}\mathbf{A}-\lambda_{k_1}\mathbf{I} &= \sum_{j=1}^{n}\lambda_{k_j}\mathbf{G}_{k_j} - \sum_{j=1}^{n}\lambda_{k_1}\mathbf{G}_{k_j}\\ &= \sum_{j=2}^{n}(\lambda_{k_j}-\lambda_{k_1})\mathbf{G}_{k_j}\\ &= \sum_{j=2}^{n}\prod_{i=1}^{1}(\lambda_{k_j}-\lambda_{k_i})\mathbf{G}_{k_j}.\end{aligned}$$

Hence, (29) is true for $p=2$. Assume that (29) is true for p, then

$$\begin{aligned}&\tilde{\mathbf{B}}_{p+1}\\ &= \prod_{i=1}^{p}(\mathbf{A}-\lambda_{k_i}\mathbf{I})\\ &= (\mathbf{A}-\lambda_{k_p}\mathbf{I})\tilde{\mathbf{B}}_p\\ &= (\mathbf{A}-\lambda_{k_p}\mathbf{I})\sum_{j=p}^{n}\prod_{i=1}^{p-1}(\lambda_{k_j}-\lambda_{k_i})\mathbf{G}_{k_j}\\ &= (\mathbf{A}-\lambda_{k_p}\mathbf{I})\left[\prod_{i=1}^{p-1}(\lambda_{k_p}-\lambda_{k_i})\mathbf{G}_{k_p}\right.\\ &\quad\left.+\sum_{j=p+1}^{n}\prod_{i=1}^{p-1}(\lambda_{k_j}-\lambda_{k_i})\mathbf{G}_{k_j}\right]\end{aligned}$$

$$
\begin{aligned}
&= \prod_{i=1}^{p-1}(\lambda_{k_p}-\lambda_{k_i})(\mathbf{A}\mathbf{G}_{k_p}-\lambda_{k_p}\mathbf{G}_{k_p}) \\
&+ \sum_{j=p+1}^{n}\prod_{i=1}^{p-1}(\lambda_{k_j}-\lambda_{k_i})(\mathbf{A}\mathbf{G}_{k_j}-\lambda_{k_p}\mathbf{G}_{k_j}) \\
&= \sum_{j=p+1}^{n}\prod_{i=1}^{p-1}(\lambda_{k_j}-\lambda_{k_i})(\mathbf{A}\mathbf{G}_{k_j}-\lambda_{k_p}\mathbf{G}_{k_j}) \\
&= \sum_{j=p+1}^{n}\prod_{i=1}^{p}(\lambda_{k_j}-\lambda_{k_i})\mathbf{G}_{k_j}
\end{aligned}
$$

where we have used the fact that, $\mathbf{A}\mathbf{G}_{k_j} = \lambda_{k_j}\mathbf{G}_{k_j}$. ■

Thus from Lemma 1, $0 \in \sigma(\tilde{\mathbf{B}}_p)$ and has multiplicity $p-1$ and convergences to the mode $(\lambda_{k_p}, \mathbf{u}_{k_p})$, which is guaranteed, provided $\prod_{i=1}^{p-1}(\lambda_{k_p}-\lambda_{k_i})$ is dominant. Furthermore, there is no need to evaluate $\tilde{\mathbf{B}}_p$ explicitly, thus avoiding matrix - matrix multiplications as the iterates $\tilde{\mathbf{B}}_p\mathbf{x}_0$ may be computed as a sequence of matrix - vector multiplications. Algorithm 5 briefly indicates the procedure to be used to **cyclically** determine the eigenvalues.

Algorithm 5: **cyclic**

call **power**($\mathbf{A}$)
$\mathrm{L}[1] = \lambda$
$\mathbf{A} = \mathbf{A} - \mathrm{L}[1]\mathbf{I}$
call **power**($\mathbf{A}$)
$\mathrm{L}[2] = \lambda$
for $k = 2$ to $N-1$ **do**
  $\mathbf{A} = \mathbf{A}(\mathbf{A} - \mathrm{L}[i]\mathbf{I})$
  call **power**($\mathbf{A}$)
  $\mathrm{L}[i+1] = \lambda$
end for
output vector $\mathbf{L}$

It is obvious that the first two modes determined are λ_1 and λ_N. Thereafter the modes are determined according to the dominant eigenvalues of the matrix $\tilde{\mathbf{B}}_p$ and are stored in the vector $\mathbf{L}$.

Example 7: Consider the matrix

$$
\mathbf{A} = \begin{bmatrix} 9 & 4 & 3 & 2 & 1 \\ 4 & 10 & 0 & 4 & 3 \\ 3 & 0 & 11 & 6 & 5 \\ 2 & 4 & 6 & 12 & 7 \\ 1 & 3 & 5 & 7 & 13 \end{bmatrix},
$$

where $\sigma(\mathbf{A}) = \{26.406875, 11.513724, 8.848950, 5.327046, 2.903405\}$. Table I indicates the results using Algorithm 5 to determine all modes. We have used $\varepsilon = 10^{-6}$ as tolerance for eigenvector convergence in **power**($\mathbf{A}$).

TABLE I
ERRORS AND MODES FOR EXAMPLE 7

Iterations	Mode	Error
2	1	1.7×10^{-10}
72	5	1.5×10^{-10}
45	2	5.3×10^{-15}
112	4	4.7×10^{-11}
17	3	4.4×10^{-12}

VII. ACCELERATING CONVERGENCE

From (10), we have

$$
\mathbf{x}_k = \frac{\lambda_1^k\mathbf{G}_1\mathbf{x}_0 + \sum_2^n \lambda_i^k\mathbf{G}_i\mathbf{x}_0}{\left(\lambda_1^{2k}\|\mathbf{G}_1\mathbf{x}_0\|_2^2 + \sum_2^n \lambda_i^{2k}\|\mathbf{G}_i\mathbf{x}_0\|_2^2\right)^{\frac{1}{2}}} \tag{30}
$$

$$
= \frac{\left(\frac{\lambda_1}{|\lambda_1|}\right)^k \frac{\mathbf{G}_1\mathbf{x}_0}{\|\mathbf{G}_1\mathbf{x}_0\|_2} + \sum_{i=2}^n \left(\frac{\lambda_i}{|\lambda_1|}\right)^k \frac{\mathbf{G}_i\mathbf{x}_0}{\|\mathbf{G}_1\mathbf{x}_0\|_2}}{\left(1 + \sum_{i=2}^n \left(\frac{\lambda_i}{\lambda_1}\right)^{2k} \frac{\|\mathbf{G}_i\mathbf{x}_0\|_2^2}{\|\mathbf{G}_1\mathbf{x}_0\|_2^2}\right)^{\frac{1}{2}}}.
$$

It follows from (30), after a binomial expansion, we may write

$$
\mathbf{x}_k = \mathbf{u}_1 + \beta^k\mathbf{C} + \mathbf{g}_k, \tag{31}
$$

where $\mathbf{u}_1 = \left(\frac{\lambda_1}{|\lambda_1|}\right)^k \frac{\mathbf{G}_1\mathbf{x}_0}{\|\mathbf{G}_1\mathbf{x}_0\|_2}$, $\beta = \frac{\lambda_2}{|\lambda_1|}$, $\mathbf{C} = \frac{\mathbf{G}_2\mathbf{x}_0}{\|\mathbf{G}_1\mathbf{x}_0\|_2}$ is an eigenvector corresponding to λ_2 and $\mathbf{g}_k$ is the appropriate error vector. Furthermore, $\beta^k \to 0$ and $\mathbf{g}_k \to \mathbf{0}$ as $k \to \infty$. Aitken acceleration [2] is one of the simplest procedures to improve the convergence of linearly converging sequences. Define the forward difference operator Δ by

$$
\Delta\mathbf{x}_k = \mathbf{x}_{k+1} - \mathbf{x}_k.
$$

Theorem 5: The Aitken iterates $\hat{\mathbf{x}}_k$ given by

$$
\hat{\mathbf{x}}_k = \mathbf{x}_k - \frac{(\Delta\mathbf{x}_k)^2}{\Delta^2\mathbf{x}_k}, \tag{32}
$$

converge faster to $\mathbf{u}_1$ than $\mathbf{x}_k$. Note that in (32) and what is to follow, multiplication and division of vectors, refers to componentwise operations.

Proof: From (31) and (32), it follows that

$$
\begin{aligned}
&\frac{\hat{\mathbf{x}}_k - \mathbf{u}_1}{\mathbf{x}_k - \mathbf{u}_1} \qquad (33)\\
&= \mathbf{1} - \frac{(\mathbf{x}_{k+1}-\mathbf{x}_k)^2}{(\mathbf{x}_{k+2}-2\mathbf{x}_{k+1}+\mathbf{x}_k)(\mathbf{x}_k-\mathbf{u}_1)} \\
&= \mathbf{1} - \frac{\beta^{2k}(\beta-1)^2\mathbf{C}^2 + 2\beta^k(\beta-1)\mathbf{C}\Delta\mathbf{g}_k + (\Delta\mathbf{g}_k)^2}{(\beta^k(\beta-1)^2\mathbf{C} + \Delta^2\mathbf{g}_k)(\beta^k\mathbf{C} + \mathbf{g}_k)} \\
&= \frac{\beta^k(\beta-1)^2\mathbf{C}\mathbf{g}_k + \beta^k\mathbf{C}\Delta^2\mathbf{g}_k + \mathbf{g}_k\Delta^2\mathbf{g}_k - 2\beta^k(\beta-1)\mathbf{C}\Delta\mathbf{g}_k - (\Delta\mathbf{g}_k)^2}{\beta^{2k}(\beta-1)^2\mathbf{C}^2 + \beta^k(\beta-1)^2\mathbf{C}\mathbf{g}_k + \beta^k\mathbf{C}\Delta^2\mathbf{g}_k + \mathbf{g}_k\Delta^2\mathbf{g}_k} \\
&= \frac{(\beta-1)^2\mathbf{C}\mathbf{g}_k - 2(\beta-1)\mathbf{C}\Delta\mathbf{g}_k + \mathbf{C}\Delta^2\mathbf{g}_k + \frac{\mathbf{g}_k\Delta^2\mathbf{g}_k}{\beta^k} - \frac{(\Delta\mathbf{g}_k)^2}{\beta^k}}{\beta^k(\beta-1)^2\mathbf{C}^2 + (\beta-1)^2\mathbf{C}\mathbf{g}_k + \mathbf{C}\Delta^2\mathbf{g}_k + \frac{\mathbf{g}_k\Delta^2}{\beta^k}\mathbf{g}_k}.
\end{aligned}
$$

Since $\mathbf{g}_k$, $\beta^k \to 0$ as $k \to \infty$, we get from (33)

$$
\begin{aligned}
&\lim_{k\to\infty} \frac{|\hat{\mathbf{x}}_k - \mathbf{u}_1|}{|\mathbf{x}_k - \mathbf{u}_1|} \\
&= \lim_{k\to\infty} \left| \frac{\frac{\mathbf{g}_k\Delta^2\mathbf{g}_k}{\beta^k} - \frac{(\Delta\mathbf{g}_k)^2}{\beta^k}}{\frac{\mathbf{g}_k\Delta^2\mathbf{g}_k}{\beta^k}} \right| \\
&= \lim_{k\to\infty} \left| \frac{\mathbf{g}_k\Delta^2\mathbf{g}_k - (\Delta\mathbf{g}_k)^2}{\mathbf{g}_k\Delta^2\mathbf{g}_k} \right| \\
&= \lim_{k\to\infty} \left| \frac{\mathbf{g}_k\mathbf{g}_{k+2} - \mathbf{g}_{k+1}^2}{\mathbf{g}_k\mathbf{g}_{k+2} - 2\mathbf{g}_k\mathbf{g}_{k+1} + \mathbf{g}_k^2} \right|. \qquad (34)
\end{aligned}
$$

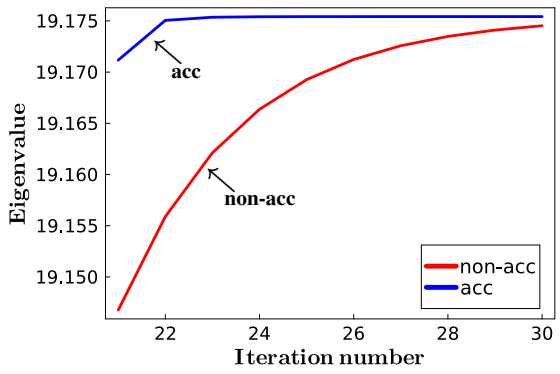

Fig. 6. Aitken acceleration for Example 7

We now assume that $\mathbf{g}_k \to \mathbf{0}$ monotonically and linearly at a rate δ, so that we may write for k large

$$\mathbf{g_{k+1}} = \delta \mathbf{g}_k.$$

Then (34) becomes

$$\lim_{k\to\infty} \frac{|\hat{\mathbf{x}}_k - \mathbf{u}_1|}{|\mathbf{x}_k - \mathbf{u}_1|} = \lim_{k\to\infty} \left| \frac{\delta^2 \mathbf{g}_k^2 - \delta^2 \mathbf{g}_k^2}{\delta^2 \mathbf{g}_k^2 - 2\delta \mathbf{g}_k^2 + \mathbf{g}_k^2} \right|$$

$$= \mathbf{0}.$$

Then $\hat{\mathbf{x}}_k \to \mathbf{u}_1$ at least super-linearly. ■

Example 8: For the matrix of Example 4 the rate of convergence to the first mode is $|\frac{\lambda_2}{\lambda_1}| \approx 0.824$. Aitken acceleration is applied to the power method iterates to speed up the convergence. This is depicted in Fig. 6 where the accelerated version has already converged by the 21_{st} iterate, shown in blue, as compared to the non accelerated version.

Let $f(z)$ be a function defined on $\sigma(\mathbf{A})$. Since, $\mathbf{A}$ is diagonalizable, it follows from the spectral theorem that

$$f(\mathbf{A}) = \sum_{i=1}^{N} f(\lambda_i)\mathbf{G}_i.$$

Now, suppose that $f(z)$ is increasing on $\sigma(\mathbf{A})$, with $\frac{f(\lambda_2)}{f(\lambda_1)} < \frac{\lambda_2}{\lambda_1}$ and that $\mathbf{A}$ is a symmetric positive definite matrix, then $f(\lambda_i) > f(\lambda_{i+1})$, $i = 1, 2, \cdots, N-1$. Thus the power method with $f(\mathbf{A})$ converges to the mode $(f(\lambda_1), \mathbf{u}_1)$. The asymptotic error constant is given by $\frac{f(\lambda_2)}{f(\lambda_1)}$. Thus λ_1 is easily determined.

Example 9: Consider $f(z) = z^p$, $p \in \mathbb{N}$, $p \geq 2$, then the asymptotic error constant is $\left(\frac{\lambda_2}{\lambda_1}\right)^p < \frac{\lambda_2}{\lambda_1}$. Thus convergence is faster.

However, evaluating $f(\mathbf{A}) = \mathbf{A}^p$ requires $(p-1)\mathcal{O}(n^3)$ multiplications [4] and is not practical for n large. When $\frac{\lambda_2}{\lambda_1}$ is close to unity, then it may be acceptable to use $p = 2$ to achieve faster convergence, due to a larger separation of the eigenvalues.

The modified power iteration [9] has the potential to generate both the first and the second eigenmodes, with the potential to decrease the asymptotic error constant to $\frac{|\lambda_3|}{|\lambda_1|}$. This is particularly useful when $\frac{|\lambda_2|}{|\lambda_1|} \approx 1$ and $|\lambda_2| >> |\lambda_3|$. However, the method can be unstable and converge to the first eigenmode only, despite the extra numerical effort.

VIII. Conclusion

We have presented some aspects of the power method, including some variations. In particular, we have shown that deflation is stable by examining the actual deflation process from a numerical analysis point of view. Furthermore, we have presented a way of avoiding the standard deflation process, which yields the same results as deflation. The knowledge of projection operators have been shown to be invaluable, in applying the power method, for cetain distributions of the spectrum of the matrix. A brief discussion of accelerating convergence is also presented. Our presentation is in detail, though not exhaustive. Till today the power method enjoys much interest and research from scientists.

References

[1] Zhong-Zhi Bai, Wen-Ting Wu, Galina V. Muratova, "The power method and beyond," Applied Numerical Mathematics, vol. 164, pp. 29–42, 2021.
[2] Richard L. Burden, J. Douglas Faires, "Numerical Analysis 9th edition," Brooks/Cole, Canada, vol. 87, 2010.
[3] Dmitrii Konstantinovich Faddeev, Vera Nikolaevna Faddeeva, "Computational Methods of Linear Algebra," W. H. Freeman, San Francisco, 1963.
[4] Gene H. Golub, Henk A. van der Vorst, "Eigenvalue Computation in the 20th century," Journal of Computational and Applied Mathematics, vol. 123, pp. 35–65, 2000.
[5] Robert Todd Gregory, David L. Karney, "A Collection of Matrices for Testing Computational Algorithms," Robert E. Krieger, New York, vol. 57, 1987.
[6] Chuanging Gu, Fei Xie, Ke Zhang, "A two-step matrix splitting iteration for computing the PageRank," Journal of Computational and Applied Mathematics, vol. 278, pp. 19–28, 2015.
[7] Carl D. Meyer," Matrix Analysis and Applied Linear Algebra," SIAM, Philadelphia, vol. 517, 2000.
[8] Congzhou Mike Sha, Nikolay V. Dokholyan, "Simple exponential acceleration of the power iteration algorithm," arXiv preprint arXiv:2109.10884, 2021.
[9] Bo Shi, Bojan Petrovic,"Implementation of the modified power iteration method to two-group monte Carlo eigenvalue problems," Annals of Nuclear Energy, vol. 38, pp. 781–787, 2011.
[10] Pravin Singh, Shivani Singh, Virath Singh, "New bounds for the maximal eigenvalues of positive definite matrices," International Journal of Applied Mathematics, vol. 35, pp. 685–691, 2022.
[11] Pravin Singh, Shivani Singh, Virath Singh, "Outer Bounds for the Extremal Eigenvalues of Positive Definite Matrices," IAENG International Journal of Applied Mathematics, vol. 53, no. 2, pp. 690–694, 2023.
[12] James Hardy Wilkinson, "The Algebraic Eigenvalue Problem," Clarendon Press, Oxford, 1965.
[13] Mu-Zheng Zhu, Ya-E Qi, "On the Eigenvalues Distribution of Preconditioned Block Two-by-two Matrix," IAENG International Journal of Applied Mathematics, vol. 46, no. 4, pp. 500–504, 2016.