East Asian Journal on Applied Mathematics
doi: 10.4208/eajam.160816.131016a

Vol. 7, No. 1, pp. 21-37
February 2017

A Filtered-Davidson Method for Large Symmetric Eigenvalue Problems

Cun-Qiang Miao*

Institute of Computational Mathematics and Scientific/Engineering Computing, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China.

Received 16 August 2016; Accepted (in revised version) 13 October 2016.

Abstract. For symmetric eigenvalue problems, we constructed a three-term recurrence polynomial filter by means of Chebyshev polynomials. The new filtering technique does not need to solve linear systems and only needs matrix-vector products. It is a memory conserving filtering technique for its three-term recurrence relation. As an application, we use this filtering strategy to the Davidson method and propose the filtered-Davidson method. Through choosing suitable shifts, this method can gain cubic convergence rate locally. Theory and numerical experiments show the efficiency of the new filtering technique.

AMS subject classifications: 15A18, 65F15, 65F50, 34L15, 65N25.

Key words: Symmetric eigenproblem, filtering technique, Chebyshev polynomials, Krylov subspace, Davidson-type method.

1. Introduction

The Davidson method [1] is an efficient iterative procedure for computing a few eigenvalues and the corresponding eigenvectors of the standard eigenvalue problem

$$Ax = \lambda x, \quad \text{with} \quad \|x\| = 1, \tag{1.1}$$

where $A \in \mathbb{R}^{n\times n}$ is a large sparse symmetric matrix and $\|\cdot\|$ denotes the Euclidean norm. The Davidson method performs a so-called Rayleigh-Ritz procedure [12] on an increasing subspace which is extended by adding a preconditioned residual to the current subspace. For the unpreconditioned Davidson method, it is equivalent to the Lanczos method [12, 14, 15]. It has been known as a very successful method, especially, when dealing with certain diagonally dominant matrices for using diagonal preconditioner in its original paper. Subsequently, Morgan and Scott [7,8] generalized the Davidson method to a more general

*Corresponding author. *Email address:* miaocunqiang@lsec.cc.ac.cn (C.-Q. Miao)

form. In the generalized Davidson method, they used a general preconditioner rather than a diagonal preconditioner.

In [17], Sleijpen and van der Vorst proposed a Jacobi-Davidson iteration method. In each step of the Jacobi-Davidson iteration method, a so-called correction equation needs to be solved. The Jacobi-Davidson iteration method is also a Davidson-type method, because the solution of the correction equation can be considered as a preconditioned residual vector. The coefficient matrix is a projection on the orthogonal complement of the current approximation which ensures the well-conditioned property of the correction equation when the approximate vector is near to the desired eigenvector. Furthermore, the Jacobi-Davidson iteration method can obtain cubic convergence rate locally. For more details of the Jacobi-Davidson iteration method and its convergence property, we refer to [16,17,19].

To obtain the preconditioned residual from the above discussions, we need to solve some linear systems which result in high computational costs, so as to the CPU time, especially for large problems. In [21], the authors proposed a new Chebyshev-Davidson method, in which the correction equation of the Davidson method is replaced by a Chebyshev polynomial filtering step which can amplify components of the desired eigenvector. This filtering technique can reduce the computational cost for just processing the matrix-vector products, although an indeterminate iteration step for the Chebyshev filter should be given in advance.

In this paper, we propose a new three-term recurrence polynomial filter by means of Chebyshev polynomials. This filter is located in the Krylov subspace spanned by a shifted matrix and the current approximate vector. Also, we give an estimate of the degree of the filtered polynomial. It can reduce the computational cost and conserve memory for its three-term recurrence relation. Furthermore, we give a stopping criterion resulted from the inverse iteration method [12,14] for the inner iteration step. For some suitable parameters, the new polynomial filter can reduce the iteration numbers for high convergence rate of the proposed filtered-Davidson method.

The remainder of this paper is organized as follows. In Section 2, some preliminaries for the filtering technique, Chebyshev polynomials and the Davidson method are given. In Section 3, the three-term recurrence polynomial is derived and we propose a stopping criterion which is easily verified. Furthermore, we propose the so-called filtered-Davidson method. Some details of the filtering technique are discussed in Section 4. We use some numerical experiments to demonstrate our results in Section 5 and, in the last section, we give some conclusions and remarks.

2. Preliminaries

In this paper, we use I to denote the identity matrix of suitable dimension. For a matrix $A \in \mathbb{R}^{n\times n}$, we use A^T to denote its transpose; this notation can be easily carried over to vectors. A Krylov subspace of order m associated with a matrix A and a vector $x \neq 0$ is defined by

$$K_m(A,x) = \text{span}\{x, Ax, \cdots, A^{m-1}x\}.$$

Then for each vector $y \in K_m(A,x)$, there is a polynomial $p_{m-1}(t)$ of degree less than or equal to $m-1$ such that y can be represented by $y = p_{m-1}(A)x$. In addition, we use $\mathscr{P}_m$ to denote the set of all polynomials of degree not greater than m.

Let $\{\lambda_i\}_{i=1}^n$ be the eigenvalues of the symmetric matrix $A \in \mathbb{R}^{n\times n}$ in an ascending order

$$\lambda_1 < \lambda_2 \le \lambda_3 \le \cdots \le \lambda_n \tag{2.1}$$

and $\{x_i\}_{i=1}^n$ be the corresponding orthonormal eigenvectors. In addition, we use $\lambda(A)$ to denote the spectrum of the matrix A.

Polynomial filtering technique is used to amplify the component in the desired parts of the spectrum relative to those in the undesired parts by processing the initial or the current vector using a suitable polynomial. For example, the solution of the well-known inverse iteration is $t = (A-\sigma I)^{-1}x$, with x and σ being the current Ritz approximation and a good approaching shift, respectively, which can be interpreted that the current approximation x is filtered by a rational polynomial $\varphi(t) = 1/(t-\sigma)$. This polynomial significantly magnifies the component of some desired eigenvector corresponding to the approaching shift σ. Filtering technique is a valuable tool to speed up the convergence of some methods for computing eigenvalues and their corresponding eigenvectors, such as the Davidson and the Lanczos methods, etc.

Chebyshev polynomials are widely used both in theory, when studying the convergence of the Krylov subspace methods [12, 14, 15], and in practice, as a filter to accelerate and improve the convergence. The reason that the Chebyshev polynomial is a well performance filter can be interpreted as follows. Suppose that the current approximation x expanded by the eigen-basis $x = \sum_{i=1}^n \alpha_i x_i$ is filtered by a polynomial $p_m(t)$ of degree m. Then the next filtered approximation can be represented as

$$\begin{aligned}\tilde{x} &= p_m(A)x \\ &= \sum_{i=1}^n \alpha_i p_m(\lambda_i)x_i \\ &= \alpha_1 p_m(\lambda_1)x_1 + \sum_{i=2}^n \alpha_i p_m(\lambda_i)x_i,\end{aligned}$$

and the goal is to find a polynomial $p_m(t)$ such that the maximum absolute value of $p_m(t)$ over $\lambda_i, i = 2,\cdots,n$ is smaller than that over λ_1 as far as possible. An alternative strategy is to seek a polynomial $p_m(t)$ such that $p_m(\lambda_1) = 1$ and its maximum absolute value in the interval $[a,b]$ containing eigenvalues $\{\lambda_i\}_{i=2}^n$ is the smallest possible for the unknown of all eigenvalues. Equivalently, the problem can be represented as

$$\min_{\substack{p_m\in\mathscr{P}_m \\ p_m(\lambda_1)=1}} \max_{t\in[a,b]} |p_m(t)|.$$

Thus the optimal polynomial is the scaled Chebyshev polynomial. Refer to [14] for more details. In [2, 13, 18], the authors took the Chebyshev polynomial as a filter to accelerate

the Lanczos and the Arnoldi algorithms. In addition, in [21] Zhou and Saad proposed the Chebyshev-Davidson algorithm by using Chebyshev polynomials as a filter to accelerate the Davidson-type method. The main idea of this method is to amplify the wanted parts of the desired eigenvector by proceeding the current approximation for several iteration steps through Chebyshev polynomials and use the augmentation vector to expand the projection subspace.

In the following, we first give a simple introduction of the real Chebyshev polynomial of the first kind [14] which are defined by

$$C_k(t) = \begin{cases} \cos\big(k\cos^{-1}(t)\big), & -1 \le t \le 1, \\ \cosh\big(k\cosh^{-1}(t)\big), & |t| > 1. \end{cases}$$

Note that $C_0(t) = 1$, $C_1(t) = t$, and it has an important three-term recurrence relation

$$C_{k+1}(t) = 2tC_k(t) - C_{k-1}(t). \tag{2.2}$$

For gradient-type method, such as the steepest descent (SD) method [10,11], the conjugate gradient (CG) method, or more generally, the locally optimal preconditioned conjugate gradient (LOPCG) method [5] and the Davidson method [1], the preconditioning strategy is implemented as $K^{-1}r$, where K is a preconditioner and $r = (A-\theta I)x$ is the residual vector with respect to the current approximation x. In the original Davidson method, the author used the preconditioner $K = \theta I - D$, where θ is the Rayleigh quotient of x, D is the diagonal part of the matrix A, that is, the preconditioned residual vector

$$t = (\theta I - D)^{-1}r$$

is used to expand the projection subspace V, and the Ritz pair of A with respect to the subspace V is used as the next approximation to the desired eigenpair. The process used to extract the approximation to the desired eigenpair is known as the Rayleigh-Ritz procedure. The Davidson method has been known as a very successful method, especially when dealing with certain symmetric problems in computational chemistry. But it should be admitted that this method depends quite heavily on the strong diagonal dominance of the matrix A.

The Davidson method can be described algorithmically as follows.

Method 2.1. (The Davidson Method)

1. Choose an initial approximate vector v with $\|v\| = 1$ and denote $V = [v]$.
2. For $k = 0,1,2,\cdots$, do:
   (a) form the projection matrix $H = V^TAV$, and compute the smallest eigenpair (θ,s) of the projection system $Hs = \theta s$;
   (b) compute the Ritz vector $x = Vs$, and the corresponding residual vector $r = (A-\theta I)x$;
   (c) test for convergence; stop if satisfied;

(d) construct a preconditioner K, and compute the new direction $t = K^{-1}r$;

(e) orthonormalize t to V and expand $V = [V, t]$.

3. EndFor

In [17], Sleijpen and van der Vorst proposed a Jacobi-Davidson iteration method, in each step of this method, a so-called correction equation

$$(I - xx^T)(A - \theta I)(I - xx^T)t = -r, \quad \text{for} \quad t \perp x, \tag{2.3}$$

where x is the current approximation with $\|x\| = 1$, $\theta = x^T Ax$ and $r = (A - \theta I)x$ are the corresponding Rayleigh quotient and the residual vector, respectively, is solved to expand the projection subspace. Denote $J(\theta, x) = (I - xx^T)(A - \theta I)(I - xx^T)$ and $J^\dagger(\theta, x)$ the pseudo-inverse of operator $J(\theta, x)$. Then $t = -J^\dagger(\theta, x)r$ can also be seen as a preconditioned residual vector. It should be noticed that the operator $J(\theta, x)$ remains positive definite and well-conditioned in $\text{span}\{x\}^\perp$ thanks to the projection to the orthogonal complement of x when the current approximation x is near to the desired eigenvector. Also, the Jacobi-Davidson iteration method possesses cubic convergence rate locally.

It should be acknowledged that a linear system should be solved at each step for this preconditioning strategy in the Davidson-type methods which result in high computational costs. In this paper, we aim to seek some polynomial $p_m(t)$ satisfying certain condition to accelerate the current approximate vector x. The filtered vector $P_m(A - \sigma I)x$ is used to expand the current projection subspace. If there is a polynomial filter, only matrix-vector products need to be executed during the whole process of this method. Note that in our method, we use the polynomial to filter the current approximate vector x rather than the residual vector r, which is the main difference between the filter technique and the preconditioning technique. In fact, vector $P_m(A - \sigma I)r$ will approximate some vector inside the subspace spanned by unwanted eigenvectors. See [21] for details.

3. Polynomial Filtering Techniques

In this section, combining the superior convergence rate of the inverse iteration method with the minimal property of the Chebyshev polynomials, we construct a polynomial filter. Moreover, we derive a three-term recurrence relation of the polynomial filter similar to Chebyshev polynomials. Based on the three-term recurrence relation, the filtered vector is easily carried out in practical use.

As we know, for a given approximate vector x to the desired eigenvector and a shift σ, a shifted linear system $(A - \sigma I)t = x$ needs to be solved in each step of the inverse iteration (INVI) method [3, 6, 12, 14]. For a large sparse system, it is difficult to solve it exactly, and an alternative choice is to implement it approximately by an iterative method. That is, the solution t satisfies

$$(A - \sigma I)t = x + d, \tag{3.1}$$

where $d = (A-\sigma I)t - x$ is the residual vector of the solution t. From the convergence analysis of the inverse iteration method [3, 6], we can see that when the residual d of the linear system (3.1) satisfies $\|d\| \le \eta_0\|r\|$ for a given nonnegative constant η_0, the inverse iteration method can gain linear convergence rate locally. In particular, cubic convergence rate can be obtained for the Rayleigh quotient iteration (RQI) method [9, 20] which is a special case of the inverse iteration method.

If the linear system (3.1) is implemented by the m-step standard Krylov subspace method, as we know, the solution t is located in the following Krylov subspace:

$$K_m(A-\sigma I, x) = \operatorname{span}\left\{x, (A-\sigma I)x, (A-\sigma I)^2 x, \cdots, (A-\sigma I)^{m-1}x\right\}, \tag{3.2}$$

and there is a polynomial $p_{m-1}(t)$ of degree not greater than $m-1$ such that $t = p_{m-1}(A-\sigma I)x$. Then, the corresponding residual vector of the linear system is $d = (A-\sigma I)p_{m-1}(A-\sigma I)x - x$, and, if

$$\left\|(A-\sigma I)p_{m-1}(A-\sigma I)x - x\right\| \le \eta_0\|r\| \tag{3.3}$$

satisfies, the next approximation t to the desired eigenvector will gain high convergence rate locally. Based on the above considerations, we aim to find a polynomial $p_{m-1}(t)$ satisfying condition (3.3) and use this polynomial as a filter to accelerate the current approximation.

Assume that we get an approximate vector x with norm unity to the desired eigenvector x_1, and σ is a given shift which is a lower bound of λ_1. From the following relation

$$\begin{aligned}\left\|(A-\sigma I)p_{m-1}(A-\sigma I)x - x\right\| &\le \left\|I-(A-\sigma I)p_{m-1}(A-\sigma I)\right\|\|x\| \\ &= \max_{\sigma_i\in\lambda(A-\sigma I)}\left|1-\sigma_i p_{m-1}(\sigma_i)\right| \\ &\le \max_{t\in[a,b]}\left|1-tp_{m-1}(t)\right|, \end{aligned}\tag{3.4}$$

where $[a, b]$ is an interval containing the spectrum $\lambda(A-\sigma I)$ of the shift matrix $A-\sigma I$ with $0 < a < b$, and $\sigma_i = \lambda_i - \sigma$, we can see that a good polynomial $p_{m-1}(t)$ would be one such that

$$\max_{t\in[a,b]}\left|1-tp_{m-1}(t)\right|$$

is minimal over all polynomials of degree $\le m-1$. It is obviously that the best such polynomial is such that $1-tp_{m-1}(t)$ is an appropriately scaled and shifted Chebyshev polynomial of degree m of the first kind, that is,

$$1-tp_{m-1}(t) = \frac{C_m\left(1+2\frac{a-t}{b-a}\right)}{C_m\left(1+2\frac{a}{b-a}\right)}. \tag{3.5}$$

Denote $\mu = (b+a)/2$, $\nu = (b-a)/2$, the above equation (3.5) can be simplified as

$$1-tp_{m-1}(t) = \frac{C_m\left(\frac{\mu-t}{\nu}\right)}{C_m\left(\frac{\mu}{\nu}\right)}. \tag{3.6}$$

Thus, the polynomial $p_{m-1}(t)$ can be given explicitly as

$$p_{m-1}(t) = \frac{1}{t}\left(1 - \frac{C_m\left(\frac{\mu-t}{\nu}\right)}{C_m\left(\frac{\mu}{\nu}\right)}\right). \tag{3.7}$$

It should be noticed that it is unwise for us to use the explicit expression of polynomial $p_{m-1}(t)$ to construct the polynomial filter, because at last we need to compute an inverse of the shift matrix $A-\sigma I$. In fact, there is no need to use the explicitly expression of the polynomial $p_{m-1}(t)$, and an alternative strategy is to derive a three-term recurrence relation similar to the Chebyshev polynomials. In the next, we derive the three-term recurrence relation of the polynomial $p_{m-1}(t)$.

Based on the Chebyshev relation (2.2) and equation (3.6), by straightforward computations, we have

$$p_0(t) = \frac{1}{\mu}, \qquad p_1(t) = \frac{4\mu - 2t}{2\mu^2 - \nu^2}. \tag{3.8}$$

In addition, using the three-term recurrence relation of the Chebyshev polynomials, we have

$$\begin{aligned}
&1 - tp_m(t)\\
=&\frac{C_{m+1}(\frac{\mu-t}{\nu})}{C_{m+1}(\frac{\mu}{\nu})}\\
=&\frac{2\frac{\mu-t}{\nu}C_m(\frac{\mu-t}{\nu}) - C_{m-1}(\frac{\mu-t}{\nu})}{2\frac{\mu}{\nu}C_m(\frac{\mu}{\nu}) - C_{m-1}(\frac{\mu}{\nu})}\\
=&\frac{2\frac{\mu-t}{\nu}\big(1-tp_{m-1}(t)\big)C_m(\frac{\mu}{\nu}) - \big(1-tp_{m-2}(t)\big)C_{m-1}(\frac{\mu}{\nu})}{2\frac{\mu}{\nu}C_m(\frac{\mu}{\nu}) - C_{m-1}(\frac{\mu}{\nu})}\\
=&\frac{2\frac{\mu}{\nu}\big(1-tp_{m-1}(t)\big)C_m(\frac{\mu}{\nu}) - 2\frac{t}{\nu}\big(1-tp_{m-1}(t)\big)C_m(\frac{\mu}{\nu}) - \big(1-tp_{m-2}(t)\big)C_{m-1}(\frac{\mu}{\nu})}{2\frac{\mu}{\nu}C_m(\frac{\mu}{\nu}) - C_{m-1}(\frac{\mu}{\nu})}\\
=&\frac{2\frac{\mu}{\nu}C_m(\frac{\mu}{\nu}) - 2\frac{\mu}{\nu}tp_{m-1}(t)C_m(\frac{\mu}{\nu}) - 2\frac{t}{\nu}\big(1-tp_{m-1}(t)\big)C_m(\frac{\mu}{\nu}) - \big(1-tp_{m-2}(t)\big)C_{m-1}(\frac{\mu}{\nu})}{2\frac{\mu}{\nu}C_m(\frac{\mu}{\nu}) - C_{m-1}(\frac{\mu}{\nu})}\\
=&1 - \frac{2\frac{\mu}{\nu}tp_{m-1}(t)C_m(\frac{\mu}{\nu}) + 2\frac{t}{\nu}\big(1-tp_{m-1}(t)\big)C_m(\frac{\mu}{\nu}) - tp_{m-2}(t)C_{m-1}(\frac{\mu}{\nu})}{2\frac{\mu}{\nu}C_m(\frac{\mu}{\nu}) - C_{m-1}(\frac{\mu}{\nu})},
\end{aligned}$$

or equivalently,

$$\begin{aligned}
p_m(t) =& \frac{2\frac{\mu}{\nu}p_{m-1}(t)C_m(\frac{\mu}{\nu}) + \frac{2}{\nu}\big(1-tp_{m-1}(t)\big)C_m(\frac{\mu}{\nu}) - p_{m-2}(t)C_{m-1}(\frac{\mu}{\nu})}{2\frac{\mu}{\nu}C_m(\frac{\mu}{\nu}) - C_{m-1}(\frac{\mu}{\nu})}\\
=& \frac{\frac{2}{\nu}C_m(\frac{\mu}{\nu}) + 2\frac{\mu-t}{\nu}p_{m-1}(t)C_m(\frac{\mu}{\nu}) - p_{m-2}(t)C_{m-1}(\frac{\mu}{\nu})}{2\frac{\mu}{\nu}C_m(\frac{\mu}{\nu}) - C_{m-1}(\frac{\mu}{\nu})}.
\end{aligned}$$

Denote $\rho_m = \frac{C_{m-1}(\mu/\nu)}{C_m(\mu/\nu)}$, then $p_m(t)$ can be simplified as

$$p_m(t) = \frac{1}{2\frac{\mu}{\nu} - \rho_m}\left(2\frac{\mu - t}{\nu}p_{m-1}(t) - \rho_m p_{m-2}(t) + \frac{2}{\nu}\right), \quad m = 1, 2, \cdots, \tag{3.9}$$

where $p_{-1}(t) = 0$. So, from the recurrence relation (3.9), we can see that the filtered vector $p_{m-1}(A-\sigma I)x$ can be obtained by the preceding two items.

For a given matrix B and a vector v, the computation of $z_m = p_m(B)v$ can be implemented algorithmically as follows.

Algorithm 3.1. (The Three-Term Recurrence Filter)

1. Given matrix B, an initial vector v, constants a and b; compute $\mu = (b+a)/2$, $\nu = (b-a)/2$ and $z_0 = (1/\mu)v$; set $z_{-1} = 0$.
2. For $m = 1, 2, \cdots$, do:
   (a) compute $C_m(\mu/\nu)$ according to relation (2.2), and compute $\rho_m = \frac{C_{m-1}(\mu/\nu)}{C_m(\mu/\nu)}$;
   (b) compute z_m as
   $$z_m = \frac{1}{2\frac{\mu}{\nu} - \rho_m}\left(\frac{2}{\nu}(\mu z_{m-1} - Bz_{m-1} + v) - \rho_m z_{m-2}\right).$$
3. EndFor

From the above discussions, we can see that the filtered vector can be constructed by carrying out Algorithm 3.1 which can be terminated once condition (3.3) satisfies. Apparently, from inequality (3.4), we see that

$$\max_{t\in[a,b]} \left|1 - tp_{m-1}(t)\right| \le \eta_0 \|r\| \tag{3.10}$$

is a sufficient condition for (3.3). We should notice that condition (3.10) gives us another point of view to explain the filtered polynomial. In fact, relation

$$\begin{aligned} \left\|I - (A-\sigma I)p_{m-1}(A-\sigma I)\right\| &= \max_{\sigma_i \in \lambda(A-\sigma I)} \left|1 - \sigma_i p_{m-1}(\sigma_i)\right| \\ &\le \max_{t\in[a,b]} \left|1 - tp_{m-1}(t)\right| \\ &\le \eta_0 \|r\| \end{aligned}$$

means that $p_{m-1}(A-\sigma I)$ can be considered as an approximation to the inverse of the shift matrix $A-\sigma I$. Therefore, the filtered vector $p_{m-1}(A-\sigma I)x$ can be seen as an approximation to the inverse iteration vector $(A-\sigma I)^{-1}x$. It gives us an interpretation of why the filtered-Davidson method can gain high convergence rate through choosing a suitable shift. In [4], Jian used this technique to construct some new preconditioners for the preconditioned steepest descent method. In the following, we present one theorem which gives an estimate of degree m of the polynomial filter based on condition (3.10).

Theorem 3.1. *If the polynomial $p_{m-1}(t)$ ($m \geq 1$) generated by Algorithm* 3.1 *satisfies condition* (3.10)*, and assumption $\eta_0\|r\| < 1$ holds, then, degree m has the following estimate*

$$m > \frac{\ln\left(\frac{1}{\eta_0\|r\|} + \sqrt{\left(\frac{1}{\eta_0\|r\|}\right)^2 - 1}\right)}{\ln \tau},$$

where $\tau = \mu/\nu + \sqrt{(\mu/\nu)^2 - 1}$, $\mu = (b+a)/2$, $\nu = (b-a)/2$, and a, b are the lower and upper bounds of the spectrum of the shift matrix $A - \sigma I$, respectively.

Proof. As

$$\min_{p \in \mathscr{P}_{m-1}} \max_{t \in [a,b]} \left|1 - t p_{m-1}(t)\right| = \frac{1}{|C_m(\frac{\mu}{\nu})|},$$

by means of the explicit expression of the Chebyshev polynomial

$$C_m(t) = \frac{1}{2}\left((t + \sqrt{t^2-1})^m + (t + \sqrt{t^2-1})^{-m}\right), \tag{3.11}$$

we have

$$\min_{p \in \mathscr{P}_{m-1}} \max_{t \in [a,b]} \left|1 - t p_{m-1}(t)\right| = \frac{2}{\left|\tau^m + \tau^{-m}\right|}.$$

Under condition (3.10), we get

$$\frac{2}{\tau^m + \tau^{-m}} \leq \eta_0\|r\|,$$

or equivalently,

$$\tau^m + \tau^{-m} \geq \frac{2}{\eta_0\|r\|}.$$

Furthermore, it holds that

$$(\tau^m)^2 - \frac{2}{\eta_0\|r\|}\tau^m + 1 \geq 0,$$

by straightforward computations, we can get

$$m > \frac{\ln\left(\frac{1}{\eta_0\|r\|} + \sqrt{\left(\frac{1}{\eta_0\|r\|}\right)^2 - 1}\right)}{\ln \tau}.$$

□

Theorem 3.1 provides another stopping criterion for terminating Algorithm 3.1 when constructing the filtered vector $p_{m-1}(A - \sigma I)x$.

In the following, we propose a new Davidson-type method accelerated by the polynomial filter which is generated by Algorithm 3.1, and we call this method as the filtered-Davidson method.

Method 3.1. (The Filtered-Davidson Method)

1. Choose an initial approximate vector v with $\|v\| = 1$ and denote $V = [v]$.
2. For $k = 0, 1, 2, \cdots$, do:
   (a) form the projection matrix $H = V^T AV$, and compute the smallest eigenpair (θ, s) of the projection system $Hs = \theta s$;
   (b) compute the Ritz vector $x = Vs$, and the corresponding residual vector $r = (A - \theta I)x$;
   (c) test for convergence, stop if satisfied;
   (d) choose a suitable shift σ, a stopping criterion η_0 and compute a, b;
   (e) call Algorithm 3.1 to construct a polynomial filter $p_{m-1}(t)$ satisfying $\|x - (A - \sigma I)p_{m-1}(A - \sigma I)x\| \leq \eta_0 \|r\|$;
   (f) let $t = p_{m-1}(A - \sigma I)x$, orthonormalize t to V and expand $V = [V, t]$.
3. EndFor

4. Detailed Discussions

In this section, we give a detailed discussion of the filtering strategy. From the above considerations, we can see that it is vitally important to choose a suitable shift σ and determine a, b which are lower and upper bounds of the spectrum of the shift matrix $A - \sigma I$ in Method 3.1.

For an approximate unit vector x with its Rayleigh quotient being $\theta = x^T Ax$, assumptions (2.1) and

$$|\lambda_1 - \theta| < \lambda_2 - \theta \tag{4.1}$$

are imposed on the following discussions. Note that assumption (4.1) can result in $\theta < (\lambda_1 + \lambda_2)/2$.

Lemma 4.1 ([14]). *Let (θ, u) be an approximate eigenpair of the symmetric matrix A with residual vector $r = (A - \theta I)u$, where u is a unit vector. Then the following estimates*

$$|\theta - \lambda_1| \leq \|r\| \quad \textit{and} \quad |\theta - \lambda_1| \leq \frac{\|r\|^2}{\lambda_2 - \theta}$$

hold under assumptions (2.1) *and* (4.1).

Next, we will give several estimates of the lower bound of λ_1 based on Lemma 4.1, which relies on $\mathcal{O}(\|r\|)$ or $\mathcal{O}(\|r\|^2)$. Through these estimates, we can provide practical ways to choose the shift σ.

On one hand, from the first inequality $|\theta-\lambda_1| \le \|r\|$ in Lemma 4.1, we know that the smallest eigenvalue $\lambda_1 \in [\theta-\|r\|, \theta+\|r\|]$, and then the lower bound of λ_1 can be obtained by

$$\lambda_1 \ge \theta - \|r\|.$$

On the other hand, the second inequality $|\theta-\lambda_1| \le \frac{\|r\|^2}{\lambda_2-\theta}$ results in $\lambda_1 \ge \theta - \frac{\|r\|^2}{\lambda_2-\theta}$. Using the fact $\lambda_2-\theta > (\lambda_2-\lambda_1)/2$ resulted from assumption (4.1), we can obtain another lower bound of λ_1 by

$$\lambda_1 > \theta - \frac{2\|r\|^2}{\lambda_2-\lambda_1}.$$

Based on the above considerations, the shift σ can be chosen as

$$\sigma = \theta - c_1\|r\| \quad \text{or} \quad \sigma = \theta - c_2\|r\|^2,$$

where $c_1 > 1$ and $c_2 > 2/(\lambda_2-\lambda_1)$, and the lower bound of the shift matrix $A-\sigma I$ can be obtained by

$$\lambda_1-\sigma \ge (c_1-1)\|r\| \quad \text{or} \quad \lambda_1-\sigma > \left(c_2 - \frac{2}{\lambda_2-\lambda_1}\right)\|r\|^2.$$

Thus, a can be chosen as $a=(c_1-1)\|r\|$ or $a=(c_2-\frac{2}{\lambda_2-\lambda_1})\|r\|^2$, correspondingly. Theorem 3.1 tells us that the lower bound of degree m decreases as parameter a increases. That is, the iteration number of Algorithm 3.1 becomes smaller and smaller as c_1 or c_2 becomes larger and larger. But we must notice that for large enough c_1 or c_2, it may make the filter less effective, because for the inverse iteration equation $(A-\sigma I)t = x$ solved exactly or inexactly, it will obtain quadratic and cubic convergence rate if the shift is chosen as $\sigma=\theta-c_1\|r\|$ and $\sigma=\theta-c_2\|r\|^2$, respectively. Furthermore, large c_1 or c_2 may waken the convergence rate of the inverse iteration equation, which reveals that if we want to make the filtered vector more effective, Algorithm 3.1 may be implemented with properly large iteration steps.

We should admit that it is impractical to choose c_2 through $c_2 > 2/(\lambda_2-\lambda_1)$, because λ_1 and λ_2 are unknown. Saad in [14] provides a practical way. For an approximation $\tilde{\lambda}_2$ to the second smallest eigenvalue λ_2, we have

$$\begin{aligned}|\lambda_2-\theta| =& |\theta-\tilde{\lambda}_2+\tilde{\lambda}_2-\lambda_2| \\ \ge& |\theta-\tilde{\lambda}_2| - |\tilde{\lambda}_2-\lambda_2| \\ \ge& |\theta-\tilde{\lambda}_2| - \|r_2\|,\end{aligned}$$

which means

$$\lambda_1 \ge \theta - \frac{\|r\|^2}{|\theta-\tilde{\lambda}_2|-\|r_2\|},$$

where $\|r_2\|$ is the residual norm of the approximation to the second smallest eigenpair. Thus, we can choose $c_2 > \frac{1}{|\theta-\tilde{\lambda}_2|-\|r_2\|}$ which is computable in practice, but we need a good approximation $\tilde{\lambda}_2$ to the second eigenvalue λ_2, in other words, the residual norm $\|r_2\|$ should be small enough to satisfy $\|r_2\| < |\theta-\tilde{\lambda}_2|$.

For the upper bound of the spectrum of the shift matrix $B = A-\sigma I$, we can use its infinite norm $b = \|B\|_\infty$ or other norms valid.

At the end of this section, we present the convergence result of the filtered-Davidson method.

Theorem 4.1. *Assume that we obtain an approximate unit vector x to the desired eigenvector x_1 by Method* 3.1*, and it admits the orthogonal direct-sum decomposition*

$$x = x_1\cos\phi + w\sin\phi, \quad \textit{with} \quad w \perp x_1, \tag{4.2}$$

where $\|w\| = 1$, ϕ is the angle between vectors x and x_1. The shift in the filtered-Davidson method is chosen as $\sigma = \theta - c\|r\|^2$, with c being a positive constant. Let $\tilde{x}$ be the next approximate vector after one step iteration and $\tilde{\phi}$ be the corresponding angle between vectors $\tilde{x}$ and x_1. For a given positive constant η_0, assume that θ is near to λ_1 such that

$$\delta = \eta_0(\lambda_n-\lambda_1)\sqrt{\frac{\theta-\lambda_1}{\lambda_2-\theta}} < 1,$$

then the following estimate holds

$$\tan\tilde{\phi} \le (\lambda_n-\lambda_1)\frac{\big(1+c(\lambda_n-\lambda_1)\big)\big(1+\eta_0(\lambda_n-\lambda_1)\big)}{(\lambda_2-\sigma)(1-\delta)\cos\phi}\sin^3\phi.$$

Proof. The filtered vector $t = p_{m-1}(A-\sigma I)x$ generated by Method 3.1 can be written as

$$(A-\sigma I)t = x + d.$$

If we adopt t to be the next approximate vector $\tilde{x}$ to the desired eigenvector, we can obtain the worst-case convergence result for the filtered-Davidson method. Also, similar to the decomposition of x in (4.2), the next approximation admits the orthogonal direct-sum decomposition

$$\tilde{x} = x_1\cos\tilde{\phi} + \tilde{w}\sin\tilde{\phi}, \quad \text{with} \quad \tilde{w} \perp x_1,$$

then, we have

$$(\lambda_1-\sigma)x_1\cos\tilde{\phi} + (A-\sigma I)\tilde{w}\sin\tilde{\phi} = x_1\cos\phi + w\sin\phi + d.$$

Multiplying both sides of the above equation from left by x_1^T and $\Pi = I-x_1x_1^T$, respectively, we obtain

$$(\lambda_1-\sigma)\cos\tilde{\phi} = \cos\phi + x_1^T d$$

and

$$\Pi(A-\sigma I)\tilde{w}\sin\tilde{\phi} = w\sin\phi + \Pi d.$$

In addition, by making use of the decomposition of x in (4.2), we can straightforwardly obtain

$$\lambda_1 - \theta = (\lambda_1 - w^T A w)\sin^2\phi, \qquad \|r\| \le (\lambda_n - \lambda_1)\sin\phi$$

and

$$\tan^2\phi = \frac{\theta-\lambda_1}{w^T A w - \theta}.$$

Hence, it holds that

$$\begin{aligned}
\tan\tilde{\phi} =& \frac{|\lambda_1-\sigma|}{\|\Pi(A-\sigma I)\tilde{w}\|}\frac{\|w\sin\phi+\Pi d\|}{|\cos\phi + x_1^T d|} \\
\le& \frac{|\lambda_1-\theta| + c\|r\|^2}{\lambda_2-\sigma}\frac{\left\|w+\frac{\Pi d}{\sin\phi}\right\|}{\left|1+\frac{x_1^T d}{\cos\phi}\right|}\tan\phi \\
\le& (\lambda_n-\lambda_1)\frac{1+c(\lambda_n-\lambda_1)}{(\lambda_2-\sigma)\cos\phi}\frac{\left\|w+\frac{\Pi d}{\sin\phi}\right\|}{\left|1+\frac{x_1^T d}{\cos\phi}\right|}\sin^3\phi. \qquad (4.3)
\end{aligned}$$

Moreover, based on condition $\|d\| \le \eta_0\|r\|$ and estimate $\tan^2\phi \le \frac{\theta-\lambda_1}{\lambda_2-\theta}$, we have

$$\frac{\|\Pi d\|}{\sin\phi} \le \eta_0(\lambda_n-\lambda_1) \quad \text{and} \quad \left|\frac{x_1^T d}{\cos\phi}\right| \le \frac{\|d\|}{|\cos\phi|} \le \eta_0(\lambda_n-\lambda_1)\sqrt{\frac{\theta-\lambda_1}{\lambda_2-\theta}}.$$

Substituting the two estimates into the inequality (4.3) leads to

$$\tan\tilde{\phi} \le (\lambda_n-\lambda_1)\frac{\big(1+c(\lambda_n-\lambda_1)\big)\big(1+\eta_0(\lambda_n-\lambda_1)\big)}{(\lambda_2-\sigma)(1-\delta)\cos\phi}\sin^3\phi.$$

□

5. Numerical Experiments

In this section, we use examples to examine the numerical behavior of the *filtered-Davidson* (**FD**) method and compare it with the *Davidson* method (**D**) [1], the *Jacobi-Davidson* (**JD**) iteration method [17] and the *Chebyshev-Davidson* (**CD**) method [21]. All runs are started from random vectors. All iteration processes are terminated once their residual norms at the current iteration step satisfying the stopping criterion $\frac{\|r^{(k)}\|}{\|r^{(0)}\|} \le 10^{-6}$,

with $r^{(k)} = (A-\theta^{(k)}I)x^{(k)}$ being the residual corresponding to the k-th approximation $x^{(k)}$ of the eigenvector x_1 and $\theta^{(k)}$ being the corresponding Rayleigh quotient.

We performed all experiments using MATLAB (with version R2013a) on a personal computer with 3.60 GHz central processing unit (Intel (R) Core (TM) i7-4790), 8.00 GB memory and Windows 10 operating system (2015).

For the Jacobi-Davidson iteration method, at each step we solve the correction equation (2.3) iteratively by making use of the *minimal residual* (**MINRES**) method preconditioned by the matrix

$$P = \left(I - x^{(k)}(x^{(k)})^T\right) K \left(I - x^{(k)}(x^{(k)})^T\right),$$

where $K = A - jI$, and j is an estimate generated by the 10-step Lanczos process. For more details of the preconditioning technique, refer to [17]. We adopt

$$\frac{\left\|\left(I - x^{(k)}(x^{(k)})^T\right)(A-\theta^{(k)}I)\left(I - x^{(k)}(x^{(k)})^T\right)t + r^{(k)}\right\|}{\|r^{(k)}\|} \leq 0.01$$

as the stopping criterion for the preconditioned MINRES method.

In the filtered-Davidson method, we choose the shift $\sigma = \theta - \|r\|^2$. The lower and upper bounds of the shift matrix $A-\sigma I$ are chosen as $a = \min\{\|r\|, \|r\|^2\}$ and $b = \|A-\sigma I\|_\infty$, respectively. When the filtered-Davidson method is implemented to compare with the Jacobi-Davidson iteration method, we adopt

$$\left\|(A-\sigma I)p_{m-1}(A-\sigma I)x - x\right\| \leq 0.1 \tag{5.1}$$

as the stopping criterion for the inner iteration of Method 3.1 (step 2(e)), but when comparing with the Chebyshev-Davidson method, for the sake of fairness, we both take ten steps for the inner polynomial iteration in these two methods instead of using the stopping criterion (5.1).

Example 5.1 ([13]). Consider the following two-dimensional partial differential equation

$$-\frac{\partial}{\partial x}\left(a(x,y)\frac{\partial}{\partial x}\right) - \frac{\partial}{\partial y}\left(b(x,y)\frac{\partial}{\partial y}\right) = \lambda u, \tag{5.2}$$

with homogeneous Dirichlet boundary conditions on the domain of the unit square. Discretizing problem (5.2) by five-point finite difference approximation on an $m \times m$ grid with the mesh size being both equal to $h = 1/(m+1)$, we obtain the standard eigenvalue problem (1.1). Note that now $n = m^2$.

In Table 1 and Table 2 we list the numbers of iteration steps (IT) and the CPU times (CPU) of the Davidson method, the Jacobi-Davidson iteration method, and the filtered-Davidson method of Example 5.1. From these tables we can see that the filtered-Davidson method outperforms the Jacobi-Davidson iteration method in terms of both number of iteration steps and computing time. As the order of the tested matrices increases, the

Table 1: Numerical Results for Example 5.1, $a(x,y)=b(x,y)=e^{-(x^2+y^2)}$.

m	D		JD		FD	
	IT	CPU	IT	CPU	IT	CPU
2^4	65	0.001	10	0.063	6	0.060
2^5	160	0.050	11	0.344	7	0.147
2^6	400	0.364	12	2.657	8	0.292
2^7	1129	4.517	13	13.163	9	0.708

Table 2: Numerical Results for Example 5.1, $a(x,y)=b(x,y)=\frac{e^{(x+y)}}{x+y}$.

m	D		JD		FD	
	IT	CPU	IT	CPU	IT	CPU
2^4	68	0.010	8	0.063	6	0.043
2^5	144	0.046	10	0.344	7	0.147
2^6	396	0.364	10	2.515	8	0.291
2^7	986	3.916	17	13.263	10	1.527

Table 3: Numerical Results for Example 5.1, $a(x,y)=b(x,y)=-e^{xy}$.

k	$m=2^6$		$m=2^7$	
	CD	FD	CD	FD
1	1.133E+04	1.131E+04	4.558E+04	4.466E+04
2	1.262E+04	7.600E+03	4.740E+04	3.344E+04
3	3.755E+03	2.765E+03	1.491E+04	1.225E+04
4	9.406E+02	5.590E+02	7.378E+03	4.619E+03
5	2.193E+02	1.402E+01	1.876E+03	5.771E+02
6	9.296E+00	1.714E-01	2.465E+03	1.144E+01
7	2.560E-01	2.628E-03	5.728E+03	1.145E-01
8	1.422E-02		2.389E+02	3.155E-03
9	2.224E-03		1.662E+01	
10			1.560E+00	
11			4.998E-01	
12			4.686E-02	
13			4.333E-03	

filtered-Davidson method also outperforms the Davidson method in terms of both number of iteration steps and computing time. These observations verify our findings that the filtered-Davidson method is time saving and converges fast with suitable shift.

In Table 3, we list the residual norms of the Chebyshev-Davidson method and the filtered-Davidson method for different iteration steps k. From this table, we can see that the filtered-Davidson method outperforms the Chebyshev-Davidson method in terms of the number of iteration steps when the degrees of the two kinds of polynomials in the two

methods are the same. Moreover, from this table we can see that the filtered-Davidson method can gain cubic convergence rate locally when the shift is chosen as $\sigma = \theta - \|r\|^2$ which is consistent with the theoretical result.

6. Concluding Remarks

In this paper, a polynomial filter which is based on a three-term recurrence relation is derived and the filtered-Davidson method is proposed. The filtered-Davidson method does not need to solve linear systems and only need matrix-vector products so that the computational cost of this new method is relatively small. Moreover, several estimates of the lower bound of the smallest eigenvalue of the matrix A and some practical shifts in the inverse iteration equation are given. By choosing suitable shifts, the filtered-Davidson method can gain cubic convergence rate locally.

Acknowledgments

The author would like to thank the referees for the helpful suggestions. This research is supported by the National Natural Science Foundation for Creative Research Groups (No. 11021101), P.R. China.

References

[1] E.R. DAVIDSON, *The iterative calculation of a few of the lowest eigenvalues and corresponding eigenvectors of large real-symmetric matrices*, J. Comput. Phys., 17(1975), pp. 87–94.
[2] H.R. FANG AND Y. SAAD, *A filtered Lanczos procedure for extreme and interior eigenvalue problems*, SIAM J. Sci. Comput., 34(2012), pp. A2220–A2246.
[3] G.H. GOLUB AND Q. YE, *Inexact inverse iteration for generalized eigenvalue problems*, BIT, 40(2000), pp. 671–684.
[4] S. JIAN, *A block preconditioned steepest descent method for symmetric eigenvalue problems*, Appl. Math. Comput., 219(2013), pp. 10198–10217.
[5] A.V. KNYAZEV, *Toward the optimal preconditioned eigensolver: locally optimal block preconditioned conjugate gradient method*, SIAM J. Sci. Comput., 23(2001), pp. 517–541.
[6] Y.-L. LAI, K.-Y. LIN AND W.-W. LIN, *An inexact inverse iteration for large sparse eigenvalue problems*, Numer. Linear Algebra Appl., 4(1997), pp. 425–437.
[7] R.B. MORGAN, *Generalizations of Davidson's method for computing eigenvalues of large nonsymmetric matrices*, J. Comput. Phys., 101(1992), pp. 287–291.
[8] R.B. MORGAN AND D.S. SCOTT, *Generalizations of Davidson's method for computing eigenvalues of sparse symmetric matrices*, SIAM J. Sci. Statisst. Copmut., 7(1986), pp. 817–825.
[9] Y. NOTAY, *Convergence analysis of inexact Rayleigh quotient iteration*, SIAM J. Matrix Anal. Appl., 24(2003), pp. 627–644.
[10] E. OVTCHINNIKOV, *Cluster robustness of preconditioned gradient subspace iteration eigensolvers*, Linear Algebra Appl., 415(2006), pp. 140–166.
[11] E.E. OVTCHINNIKOV, *Sharp convergence estimates for the preconditioned steepest descent method for Hermitian eigenvalue problems*, SIAM J. Numer. Anal., 43(2006), pp. 2668–2689.
[12] B.N. PARLETT, *The Symmetric Eigenvalue Problems*, SIAM, Philadelphia, PA, 1998.

[13] Y. SAAD, *Chebyshev acceleration techniques for solving nonsymmetric eigenvalue problems*, Math. Comp., 42(1984), pp. 567–588.
[14] Y. SAAD, *Numerical Methods for Large Eigenvalue Problems, Second Edition*, SIAM, Philadelphia, PA, 2011.
[15] Y. SAAD, *On the rates of convergence of the Lanczos and the Block-Lanczos methods*, SIAM J. Numer. Anal., 17(1980), pp. 687–706.
[16] G.L.G. SLEIJPEN, A.G.L. BOOTEN, D.R. FOKKEMA AND H.A. VAN DER VORST, *Jacobi-Davidson type methods for generalized eigenproblems and polynomial eigenproblems*, BIT, 36(1996), pp. 595–633.
[17] G.L.G. SLEIJPEN AND H.A. VAN DER VORST, *A Jacobi-Davidson iteration method for linear eigenvalue problems*, SIAM J. Matrix Anal. Appl., 17(1996), pp. 401–425.
[18] D.C. SORENSEN, *Implicit application of polynomial filters in a k-step Arnoldi method*, SIAM J. Matrix Anal. Appl., 13(1992), pp. 357–385.
[19] J. VAN DEN ESHOF, *The convergence of Jacobi-Davidson iterations for Hermitian eigenproblems*, Numer. Linear Algebra Appl., 9(2002), pp. 163–179.
[20] F. XUE AND H.C. ELMAN, *Convergence analysis of iterative solvers in inexact Rayleigh quotient iteration*, SIAM J. Matrix Anal. Appl., 31(2009), pp. 877–899.
[21] Y.-K. ZHOU AND Y. SAAD, *A Chebyshev-Davidson algorithm for large symmetric eigenproblems*, SIAM J. Matrix Anal. Appl., 29(2007), pp. 954–971.