# Single-Precision Calculation of Iterative Refinement of Eigenpairs of a Real Symmetric-Definite Generalized Eigenproblem by Using a Filter Composed of a Single Resolvent

Hiroshi Murakami[1],[a]

**Abstract:** By using a filter, we calculate approximate eigenpairs of a real symmetric-definite generalized eigenproblem $A\mathbf{v} = \lambda B\mathbf{v}$ whose eigenvalues are in a specified interval. In our experiments in this paper, the IEEE-754 single-precision floating-point (binary 32bit) number system is used for all calculations.

In general, a filter is constructed by using some resolvents $\mathcal{R}(\rho)$ with different shifts $\rho$. For a given vector $\mathbf{x}$, an action of a resolvent $\mathbf{y} := \mathcal{R}(\rho)\mathbf{x}$ is given by solving a system of linear equations $C(\rho)\mathbf{y} = B\mathbf{x}$ for $\mathbf{y}$, here the coefficient $C(\rho) = A - \rho B$ is symmetric. We assume to solve this system of linear equations by matrix factorization of $C(\rho)$, for example by modified Cholesky method ($LDL^T$ decomposition method). When both matrices $A$ and $B$ are banded, $C(\rho)$ is also banded and modified Cholesky method for banded system can be used to solve the system of linear equations. The filter we used is either a polynomial of a resolvent with a real shift, or a polynomial of an imaginary part of a resolvent with an imaginary shift. We use only a single resolvent to construct the filter so to reduce both amounts of calculation to factor matrices and especially storage to hold factors of matrices. The most disadvantage when we use only a single resolvent rather than many is, such a filter cannot have good properties especially when calculation is made in single-precision computation. Therefore, approximate eigenpairs required are not in good accuracy if they are extracted from the set of vectors obtained by an application of a combination of $B$-orthonormalization and filtering to a set of initial random vectors, But we can show by experiments, approximate eigenpairs required are refined well if they are extracted from the set of vectors obtained by a few applications of a combination of $B$-orthonormalization and filtering to a set of initial random vectors.

**Keywords:** Eigenproblem, Filter, Single Resolvent, Iterative Refinement, Single-Precision, Orthonormalization

## 1. Introduction

For a given generalized eigenproblem $A\mathbf{v} = \lambda B\mathbf{v}$ whose matrices $A$ and $B$ are real-symmetric and $B$ is positive-definite, by using a filter we try to solve all approximate eigenpairs whose eigenvalues are in a specified real interval $[a, b]$. There are many references of methods which can be classified as filter diagonalization methods [1], [3], [5], [7], [8], [10], [15].

Corresponding to the eigenproblem, the filter is composed of some resolvents $\mathcal{R}(\rho_i) \equiv (A - \rho_i B)^{-1}B$ whose shift $\rho_i$ is a complex number.

For a given vector $\mathbf{x}$, an application of the resolvent $\mathbf{y} \leftarrow \mathcal{R}(\rho)\mathbf{x}$ is calculated by solving a system of linear equations $C(\rho)\mathbf{y} = B\mathbf{x}$ for $\mathbf{y}$ with the shifted matrix $C(\rho) \equiv A - \rho B$ as the coefficient matrix. In present study, we assume to solve this kind of system by some direct method which uses matrix factorization.

When the shift $\rho$ is a real number, the matrix $C(\rho)$ is real-symmetric. When $\rho$ is a real number less than the minimum eigenvalue of the eigenproblem, the matrix is real-symmetric and positive-definite. When the shift is an imaginary number, the matrix is complex-symmetric and non-singular. For a symmetric matrix either real or complex, the modified Cholesky method can be used to solve the system of linear equations by using a matrix-decomposition and then forward and backward substitutions. The modified Cholesky method for complex-symmetric matrices is derived from the one for real-symmetric matrices by just replacing numbers and arithmetic operations from real ones to complex ones.

When the problem size is very large, with limited computing resources, both amounts of computation for matrix-decompositions of shifted matrices and especially storage to hold factors of matrices tend to restrict the calculation, which are proportional to the number of resolvents used to construct the filter. So, it is desirable to reduce the number of resolvents to compose the filter. There are two types of filters which is composed of only a single resolvent: 1) The filter which is a real polynomial of a resolvent with a real shift, 2) The filter which is a real polynomial of the imaginary-part of a resolvent with an imaginary shift. In present study, we used a Chebyshev polynomial to express the "real polynomial" of the filter in order to make the filter design simple.

[1] Department of Mathematical Sciences, Tokyo Metropolitan University, Hachioji, Tokyo 192-0397, Japan

[a] mrkmhrsh@tmu.ac.jp

However, characteristics of filters of these simple types are not very good, since they are composed of a single resolvent instead of many ones, and also the real polynomial is expressed by using a Chebyshev polynomial in order to make the filter design simple. For example, their transfer functions cannot have steep changes of values, thus $\mu - 1$ the geometrical ratio of widths of transition-bands and pass-bands cannot be made so small. Also, (when the upper-bound of the transfer function magnitude in stop-bands $g_{\mathrm{S}}$ is set very small) the value of $1/g_{\mathrm{P}}$ is large which is the max-min ratio of the transfer function of the filter in the pass-band $\lambda \in [a, b]$. When this max-min ratio is very large, rates of required eigenvectors contained in the set of vectors tend to have different orders of magnitudes after an application of the filter. Therefore, within a vector, values of those eigenvectors whose magnitudes of transfer-rates are smaller lose accuracy by round-off errors since they are suppressed by those eigenvectors whose magnitudes of transfer-rates are larger. By this reason, for those eigenvectors which are extracted from a set of filtered vectors, their accuracy tend to be lower if magnitudes of transfer-rates of them are smaller. Therefore, some approximate eigenpairs may not attain the required level of accuracy.

In the above explanation about filtering method, we assumed the filter is applied only once to a set of initial vectors. The following procedure shows how to calculate approximate eigenpairs by applying the filter once.

1) Let $Y^{(0)}$ be a set of $m$ random column vectors as initial vectors.

2) $B$-orthonormalize $Y^{(0)}$ to make $X^{(1)}$;
$X^{(1)}$ is filtered to make $Y^{(1)}$.

3) Taking into account the filter's characteristics, approximate eigenpairs are constructed from both sets of vectors $X^{(1)}$ and $Y^{(1)}$.

## 2. Iterative Refinement of Eigenpairs by Using a Filter

When characteristics of the filter are not good, the accuracy of approximate eigenpairs obtained by an application of the filter started from a set of random vectors is also not good. However, in that case, approximate eigenpairs can be refined by a few iterations of a combination of $B$-orthonormalization and filtering.

The following procedure shows a method to calculate refined approximate eigenpairs by applying the filter `IT` times.

```
1) Let $Y^{(0)}$ be a set of $m$ random column vectors as initial vectors.
2) Iterate the following for $i = 1, \ldots, \mathrm{IT}$
       $B$-orthonormalize $Y^{(i-1)}$ to make $X^{(i)}$;
       $X^{(i)}$ is filtered to make $Y^{(i)}$.
3) Approximate eigenpairs are constructed from both sets of vectors $X^{(\mathtt{IT})}$ and $Y^{(\mathtt{IT})}$ considering the threshold of the transfer rate of the filter for the required eigenvectors.
```

(During the iteration in the above step 2, if the decrease of the effective rank of the set of vectors is found by $B$-orthonormalization with a threshold, then we decrease $m$ the number of vectors in the set.)

The process of orthonormalization prevents eigenvectors of small magnitudes of transfer-rates from losing their accuracy by numerical round-off errors. It prevents the set of vectors dominated by those eigenvectors whose magnitudes of transfer-rates are larger. The principle to use orthogonalization of vectors in each iteration step is called "orthogonal iteration" and it is well-known[4], [12], [13].

### 2.1 Orthonormalization by Using SVD

In our experiments, we used $B$-SVD method to make $B$-orthonormalization of a given set of vectors, which is the singular value decomposition method with matrix $B$ as the inner-product metric.

The procedure is described below.

1) Make a size $m$ real symmetric-definite matrix $G \Leftarrow Y^T BY$.

2) Make an eigenvalue decomposition $G \Rightarrow UDU^T$ by using Jacobi method[11][2] to obtain size $m$ diagonal matrix $D$ and orthogonal matrix $U$ (diagonals of $D$ are prepared in descending order).

3) We make a set of $B$-orthogonal column vectors by $Y \Leftarrow YU$.

When round-off errors in making $G$ are accumulated, or the original $Y$ is ill-conditioned, the $B$-orthogonality of the set of column vectors $Y$ which is obtained by the above procedure may not be sufficient. To improve the $B$-orthogonality, we may iterate the above steps from 1) to 3) a few times, and when $G$ becomes almost diagonal, then the $B$-orthogonality is regarded as sufficient and the iteration is terminated.

After the $B$-orthogonalization is finished, we make $B$-normalization with a threshold by the following method. We reject those columns of $Y$ whose norms (square roots of diagonals of $G$) are below a threshold, and for those columns remained we multiply reciprocals of their norms, then $Y$ will be a set of $B$-orthonormal vectors. In the new set of $Y$, the number of vectors are decreased by the number of rejected ones.

On the $B$-orthogonalization method, there are other literatures[14][6].

## 3. Filters Composed of a Single Resolvent

We solve those eigenpairs whose eigenvalues are in a specified interval $[a, b]$. We have two kinds of filters depending on the location of the interval.

**Filter to Solve Eigenpairs with Lowest Eigenvalues**

When the interval is located at the lower-end of the eigenvalue distribution, we use a filter $\mathcal{F}$ which is a polynomial of a single resolvent $\mathcal{R}(\rho)$ whose shift $\rho$ is a real number less than the minimum eigenvalue $\lambda_{\min}$, and the polynomial is represented by a degree $n$ Chebyshev polynomial $T_n(x)$ (1). Here, $g_{\mathrm{S}}$ is the tight upper-bound of the transfer function magnitude $|g(t)|$ of the filter in the stop-band, $\gamma$ is a real constant, and $I$ is the identity operator.

$$\mathcal{F} = g_{\mathrm{S}}\, T_n(2\gamma\,\mathcal{R}(\rho) - I)\,. \tag{1}$$

**Filter to Solve Eigenpairs with Internal Eigenvalues**

When the interval is located at an interior of the eigenvalue distribution (or at any position), we use a filter $\mathcal{F}$ which is a polynomial of the imaginary part of a single resolvent $\mathcal{R}(\rho')$ whose shift $\rho'$ is an imaginary number, and the polynomial is represented by a Chebyshev polynomial of degree $n$ (2). Here, $g_{\mathrm{S}}$ is the tight upper-bound of the transfer function magnitude $|g(t)|$ of the filter in stop-bands, $\gamma\,'$ is a real constant, Im is the operator to take the imaginary part of a matrix, and $I$ is the identity operator.

$$\mathcal{F} = g_{\mathrm{S}}\, T_n(2\gamma\,' \,\mathrm{Im}\,\mathcal{R}(\rho') - I)\,. \tag{2}$$

### 3.1 Designs of Filters Composed of a Single Resolvent

In this paper, we specify the filter composed of a single resolvent by using a triplet of parameter $(n, \mu, g_{\mathrm{S}})$. Here, $n$ is the degree of Chebyshev polynomial of the first kind, and $\mu$ is the position of the boundary between transition-band and stop-band of the filter (in the normalized coordinate $t$), and $g_{\mathrm{S}}$ is the tight upper-bound of the transfer function magnitude $|g(t)|$ of the filter in stop-bands. (In pass-band, the maximum and the minimum of the transfer function are 1 and $g_{\mathrm{P}}$, respectively.)

If we ignored the effect of numerical round-off errors, in each filtered vector, the ratio of unrequired eigenvectors to the ratio of required eigenvectors is reduced by a small factor $g_{\mathrm{S}}/g_{\mathrm{P}}$ or less for each application of the filter.

Below, we explicitly show how filters are constructed.

#### 3.1.1 Case to Use a Real Shift for the Resolvent

In this case, we assume the interval $[a, b]$ contains the lowest eigenvalues in the eigenvalue distribution, and $a$ is no greater than the minimum eigenvalue.

From a given triplet of parameters $(n, \mu, g_{\mathrm{S}})$, we calculate the real shift $\rho$ and the real coefficient $\gamma$ (and also $g_{\mathrm{P}}$) by following expressions in (3).

$$\begin{cases} \sigma & \leftarrow \quad \mu/\sinh\left(\frac{1}{2n}\cosh^{-1}\frac{1}{g_{\mathrm{S}}}\right), \\ \rho & \leftarrow \quad a - (b-a)\,\sigma, \\ \gamma & \leftarrow \quad (b-a)(\sigma+\mu), \\ g_{\mathrm{P}} & \leftarrow \quad g_{\mathrm{S}}\cosh\left\{2n\sinh^{-1}\sqrt{(\mu-1)/(1+\sigma)}\right\}. \end{cases} \tag{3}$$

Then the filter $\mathcal{F}$ is given by the expression (1).

The normalized coordinate $t$ of $\lambda$ in this case is the linear function $t \equiv (\lambda - a)/(b - a)$ which maps between intervals $\lambda \in [a, b]$ and $t \in [0, 1]$. Then, both transfer functions $f(\lambda)$ and $g(t)$ are given by following expressions in (4).

$$\begin{cases} f(\lambda) & = \quad g_{\mathrm{S}}\, T_n\left(2\gamma \times \dfrac{1}{\lambda-\rho} - 1\right), \\ g(t) & = \quad g_{\mathrm{S}}\, T_n\left(2 \times \dfrac{\mu+\sigma}{t+\sigma} - 1\right). \end{cases} \tag{4}$$

#### 3.1.2 Case to Use an Imaginary Shift for the Resolvent

In this case, we use an imaginary shift, then the interval $[a, b]$ can be placed anywhere. From the triplet of parameters $(n, \mu, g_{\mathrm{S}})$, by using expressions in (5), we calculate the imaginary shift $\rho'$ and the real coefficient $\gamma\,'$ (and also $g_{\mathrm{P}}$).

$$\begin{cases} \sigma & \leftarrow \quad \mu/\sinh\left(\frac{1}{2n}\cosh^{-1}\frac{1}{g_{\mathrm{S}}}\right), \\ \rho' & \leftarrow \quad \frac{a+b}{2} + \left(\frac{b-a}{2}\right) \times \sigma\sqrt{-1}, \\ \gamma\,' & \leftarrow \quad \left(\frac{b-a}{2}\right) \times \frac{\mu^2+\sigma^2}{\sigma}, \\ g_{\mathrm{P}} & \leftarrow \quad g_{\mathrm{S}}\cosh\left\{2n\sinh^{-1}\sqrt{(\mu^2-1)/(1+\sigma^2)}\right\}. \end{cases} \tag{5}$$

Then the filter $\mathcal{F}$ is given by the expression (2).

In this case, the normalized coordinate $t$ of $\lambda$ which maps between intervals $\lambda \in [a, b]$ and $t \in [-1, 1]$ is the linear function $t \equiv (2\lambda - a - b)/(b - a)$.

Then, both transfer functions $f(\lambda)$ and $g(t)$ are given by following expressions in (6).

$$\begin{cases} f(\lambda) & = \quad g_{\mathrm{S}}\, T_n\left(2\gamma\,' \,\mathrm{Im}\,\dfrac{1}{\lambda-\rho'} - 1\right), \\ g(t) & = \quad g_{\mathrm{S}}\, T_n\left(2 \times \dfrac{\mu^2+\sigma^2}{t^2+\sigma^2} - 1\right). \end{cases} \tag{6}$$

### 3.2 Application of a Simple-Type Filter to a Set of Vectors

Our simple-type filter $\mathcal{F}$, which uses a degree $n$ Chebyshev polynomial of the first kind, has the following form in expression (7).

$$\mathcal{F} = g_{\mathrm{S}}\, T_n(\mathcal{X})\,. \tag{7}$$

Here, the operator $\mathcal{X}$ is $2\gamma\mathcal{R}(\rho) - I$ if the shift $\rho$ is a real number, and it is $2\gamma'\,\mathrm{Im}\,\mathcal{R}(\rho') - I$ if the shift $\rho'$ is an imaginary number. The action of this operator $\mathcal{X}$ can be easily calculated if the action of a resolvent either $\mathcal{R}(\rho)$ or $\mathcal{R}(\rho')$ is calculated.

From a set of column vectors $V$, we define $V^{(\ell)} \equiv T_\ell(\mathcal{X})\,V$, and we compute $V^{(n)}$ from $V$ by using the three-term recursion relation (8).

$$\begin{cases} V^{(0)} & \leftarrow V, \\ V^{(1)} & \leftarrow \mathcal{X}\,V, \\ V^{(\ell)} & \leftarrow 2\,\mathcal{X}\,V^{(\ell-1)} \;-\; V^{(\ell-2)} \;\; (\ell \geq 2). \end{cases} \tag{8}$$

Then, an application of the filter $\mathcal{F}$ to a set of vectors $V$ is given by the expression (9).

$$\mathcal{F}\,V = g_{\mathrm{S}}\, V^{(n)}\,. \tag{9}$$

For a filter of degree $n$, by using this three-term recursion relation, there are $n$ applications of a resolvent to a set of $m$ vectors. Each application of a resolvent is calculated by solving a set of $m$ systems of linear equations with a common coefficient matrix. For a shift $\widetilde{\rho}$ which is either real or imaginary, the calculation of $Y \leftarrow \mathcal{R}(\widetilde{\rho})X$ is made by solving $C(\widetilde{\rho})Y = BX$ for $Y$. Here, $C(\widetilde{\rho}) \equiv A - \widetilde{\rho}\,B$, and both matrices $X$ and $Y$ consist of $m$ column vectors. We have assumed that a system of linear equations is solved by using some direct method, therefore we only have to factor the symmetric matrix $C(\widetilde{\rho})$ once, and by using the factors we solve this set of $m$ systems of equations not column by column but as a whole in order to make a good data reference locality.

There are $n$ applications of the same resolvent inside an application of the filter, however we have to factor $C(\widetilde{\rho})$ only once as long as the factors of the matrix can be hold to reuse. In the iterative refinement of eigenpairs by using a filter, the same filter is applied a few times, in that case also we have to make the matrix factorization corresponding to the filter only once in all as long as the factors can be hold.

## 4. About Present Experiments

### 4.1 Filters Used in Experiments

In experiments in present paper, the filter's parameter $\mu$ is always set to 1.5, and also the parameter $g_{\mathrm{S}}$ is set to `1E-5`. For

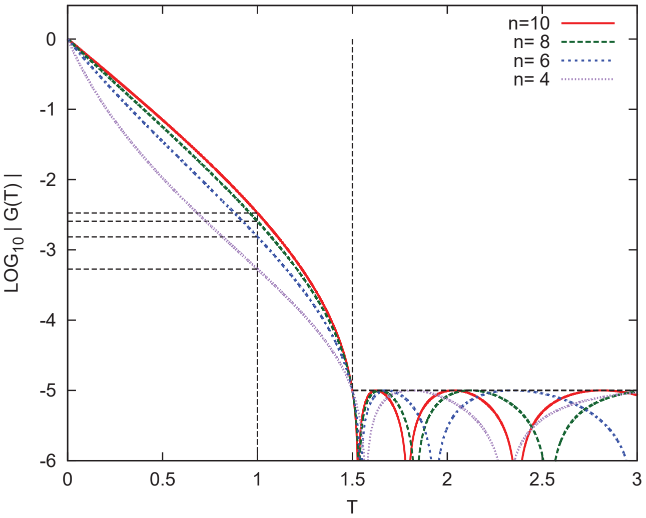

**Fig. 1** Transfer function magnitudes $|g(t)|$ of filters which consist of a single resolvent with a real shift ( $\mu = 1.5, g_S = 1\text{E}{-5}$ ).

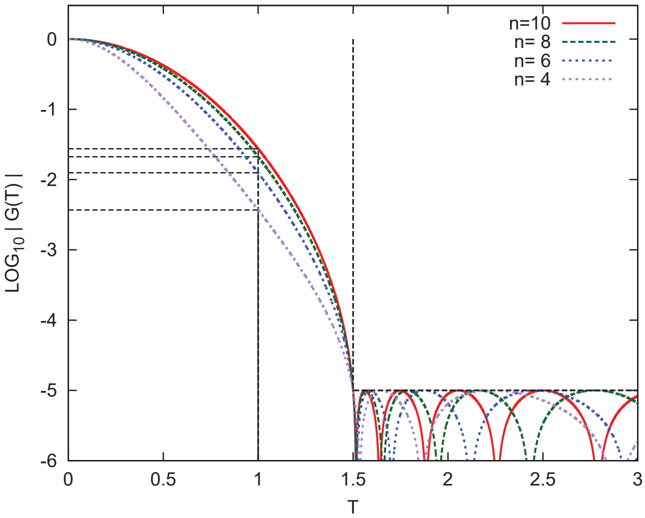

**Fig. 2** Transfer function magnitudes $|g(t)|$ of filters which consist of a single resolvent with an imaginary shift ( $\mu = 1.5, g_S = 1\text{E}{-5}$, right-half ).

each case when the interval of eigenvalues for eigenpairs is at the lower-end of the eigenvalue distribution or when the interval is interior of the distribution, we used 4 settings for filters: the degree $n$ is changed to 4, 6, 8 and 10. For these 4 settings of filters, values of $g_P$ and $g_S/g_P$ are tabulated. For filters to solve eigenpairs with lowest eigenvalues by using a real shift (**Table 1**), and for filters which solve eigenpairs with internal eigenvalues by using an imaginary shift (**Table 2**). For these filters, their transfer function magnitudes $|g(t)|$ are plotted in graphs. Here $t$ is the normalized coordinate of the eigenvalue $\lambda$ (The interval $\lambda \in [a, b]$ is mapped linearly to the interval $t \in [0, 1]$ when filters are for eigenvalues at the lower-end, and it is mapped to the interval $t \in [-1, 1]$ when filters are for internal eigenvalues). For filters which are for lowest eigenvalues, their transfer function magnitudes $|g(t)|$ are plotted in the graph (**Fig. 1**). For filters which are for internal eigenvalues, their transfer function magnitudes are plotted in the graph (**Fig. 2**).

**Table 1** Filters which consist of a single resolvent with a real shift ( $\mu = 1.5$, $g_S = 1\text{E}{-5}$ ).

| $n$ | $g_P$ | $g_S/g_P$ |
|---|---|---|
| 4 | 5.33E-4 | 1.88E-2 |
| 6 | 1.53E-3 | 6.54E-3 |
| 8 | 2.55E-3 | 3.92E-3 |
| 10 | 3.34E-3 | 2.99E-3 |

**Table 2** Filters which consist of a single resolvent with an imaginary shift ( $\mu = 1.5$, $g_S = 1\text{E}{-5}$ ).

| $n$ | $g_P$ | $g_S/g_P$ |
|---|---|---|
| 4 | 3.69E-3 | 2.71E-3 |
| 6 | 1.25E-2 | 8.01E-4 |
| 8 | 2.11E-2 | 4.73E-4 |
| 10 | 2.74E-2 | 3.65E-4 |

### 4.2 About Single-Precision Calculation

For all calculations in this paper, we used only IEEE-754 standard single-precision floating-point numbers and their operations (binary 32bit). The single-precision has about 7.22 digits of precision in decimal.

In our previous paper[9], we used IEEE-754 double-precision and quadruple-precision floating-point numbers to make experiments of iterative-refinement of eigenpairs by using a filter composed of a single resolvent. Although it is more difficult to make robust calculations with single-precision than with double-precision or quadruple-precision, recently attention has been focused on methods of reducing power consumption by performing calculations with low precision, therefore we calculated all experiments in this paper with single-precision only.

### 4.3 Generalized Eigenproblem Derived from FEM

In each experiment, a real symmetric-definite generalized eigenproblem (10) is used for example, which comes from a discretization by finite element method (FEM) of the 3-D Laplacian eigenproblem (11) in a cube whose sides have a length $\pi$ with zero-Dirichlet boundary condition.

$$A\mathbf{v} = \lambda B\mathbf{v}\,. \tag{10}$$

$$-\Delta\Psi(x, y, z) = \lambda\,\Psi(x, y, z)\,. \tag{11}$$

By the equi-division of sides in three directions into $N_1+1$, $N_2+1$, $N_3+1$ sub-intervals, the cube is partitioned into $(N_1+1)(N_2+1)(N_3+1)$ finite elements in total (we assume $N_1 \le N_2 \le N_3$). Tri-linear functions are used for basis of expansion inside each finite element. With zero-Dirichlet boundary condition, both matrices $A$ and $B$ have a size $N = N_1N_2N_3$ and a lower bandwidth $w_L = 1 + N_1 + N_1N_2$. (All eigenvalues are positive for this type of matrix generalized eigenproblem which is derived from FEM discretization of the Laplacian eigenproblem with zero-Dirichlet boundary condition.)

### 4.4 Definition of Relative Residual

In each experiment, the accuracy of an approximate eigenpair $(\lambda, \mathbf{v})$ is judged by using the relative residual $\Theta$ which is defined by the following expression (12).

$$\Theta \equiv \frac{\|A\mathbf{v} - \lambda B\mathbf{v}\|}{\|\lambda B\mathbf{v}\|}\,. \tag{12}$$

This value does not depend on the normalization of the vector $\mathbf{v}$, and the value is also unchanged when both matrices $A$ and $B$ are multiplied by a common non-zero scaling factor. In experiments, the Euclidean distance is used for the norm of a vector $\|\cdot\|$. When $\phi$ is the angle between both position vectors $A\mathbf{v}$ and $\lambda B\mathbf{v}$ in Euclidean space, then we have an inequality $\sin\phi \le \Theta$.

It is efficient to calculate relative residuals for many approximate eigenpairs together. Both $AV$ and $BV$ are calculated by matrix multiplications to reduce data references to matrices $A$ and $B$ to one each. Here $V$ is a matrix whose columns are vectors of approximate eigenpairs.

## 5. Results of Single-Precision Calculations of Iterative Refinement

We show two examples Exam-1 and Exam-2. In each example, the same FEM partitioning $(N_1, N_2, N_3) = (50, 60, 70)$ is used, therefore their equations of the eigenproblem (10) are the same. The size of both matrices $A$ and $B$ is $N = 210000$, and their lower bandwidth is $w_L = 3051$.

We solve those eigenpairs of the generalized eigenproblem whose eigenvalues are in a specified interval. The specified interval $[a, b]$ for eigenvalue is $[0, 100]$ for Exam-1, and $[100, 200]$ for Exam-2.

We iterated the combination of $B$-orthonormalization and filtering up to 6 times for Exam-1, and up to 4 times for Exam-2. In the first iteration, a set of vectors is generated from uniform random numbers, and it is $B$-orthonormalized, and the filter is applied to it.

One of the reasons why calculation of the filter diagonalization in single-precision computation is more difficult than the one in double-precision computation is as follows. In the case of calculations in single-precision computation, the proportion of required eigenvectors contained in the set of initial vectors which is generated from uniform random numbers is not so much significant compared with the level of round-off error (especially when the size of eigenproblem is large).

Another reason is that ranges of feasible filter characteristics are quite limited in single-precision computation than in double-precision computation. For example, since single-precision number has only about 7 digits of precision, a tiny value, such as $10^{-12}$ cannot be realized for the value of $g_{\mathrm{S}}$ of a filter. In the calculation of the filter application, it is not effective to set the value of $g_{\mathrm{S}}$ smaller than the level of the round-off error, because small magnitudes of filter transfer-rates in stop-bands are realized by numerical cancellations.

The computer system which we used in experiments was a single node of Oakforest-PACS system (Fujitsu PRIMERGY CX1640M1). It has a single intel Xeon Phi 7250(1.4GHz, 68 cores) for CPU with 96 GiB DDR4 memory and 16 GiB MCDRAM. The operating system was CentOS 7.6. The source code for experiments was written in Fortran 90 language with OpenMP directives. For the compiler, we used intel fortran v19.0.5.281 with compile options (”`-fast -xMIC-AVX512 -align array64byte -qopenmp`”). For all examples, the number of OpenMP threads we used was 204.

### 5.1 Exam-1: Approximate Eigenpairs with Lowest Eigenvalues

In this example Exam-1, we try to solve those eigenpairs whose eigenvalues are in the interval $[a, b] = [0, 100]$. There are 402 such eigenpairs to be solved. Since eigenvalues of these eigenpairs are at the lower-end of the eigenvalue distribution, we used a real shift for the resolvent in each filter for this example.

There are 764 eigenvalues in the interval $[a, b'] = [0, 150]$, which is the union of the pass-band and the transition-band of the filter with $\mu = 1.5$. Since it is desirable to use more initial vectors than that number, we set the number of the initial vectors to $m = 800$ for calculations.

When $g_{\mathrm{S}} = 10^{-5}$ and the degree $n$ is increased from 4 to 10 by 2 (four tables from Table 3 to Table 6, and four figures from Figure 3 to Figure 6), numbers of eigenpairs obtained are wrong for the case of `IT` = 1, but numbers of eigenpairs obtained are correct and 402 for cases `IT` is 2 or more. For cases the degree $n$ is 6 or more, even for those eigenpairs with eigenvalues near the upper-end of the interval which are slow to be improved, their relative residuals decreased well as the number of iterations increased.

In each of tables (from Table 3 to Table 6), we also show the total elapsed time consumed to obtain approximate eigenpairs by `IT` times applications of the filter combined with $B$-reorthonormalizations to a set of $m$ vectors and also including the Cholesky decomposition once to prepare the filter. The graph is also shown in Figure 7.

**Table 3** Exam-1 ($n = 4$, $g_{\mathrm{S}} = 1\mathrm{E}{-5}$, $\mu = 1.5$, $m = 800$)

| IT | # pairs | Max of Θ | Elapsed time(s) |
|---|---|---|---|
| 1 | 139 | 1.6E-01 | 164 |
| 2 | 402 | 2.7E-02 | 236 |
| 3 | 402 | 1.2E-03 | 261 |
| 4 | 402 | 3.5E-04 | 285 |
| 5 | 402 | 3.5E-04 | 322 |
| 6 | 402 | 3.5E-04 | 358 |

**Table 4** Exam-1 ($n = 6$, $g_{\mathrm{S}} = 1\mathrm{E}{-5}$, $\mu = 1.5$, $m = 800$)

| IT | # pairs | Max of Θ | Elapsed time(s) |
|---|---|---|---|
| 1 | 222 | 2.2E-01 | 191 |
| 2 | 402 | 1.0E-02 | 254 |
| 3 | 402 | 2.9E-04 | 336 |
| 4 | 402 | 2.9E-04 | 426 |
| 5 | 402 | 2.9E-04 | 473 |
| 6 | 402 | 2.9E-04 | 535 |

**Table 5** Exam-1 ($n = 8$, $g_{\mathrm{S}} = 1\mathrm{E}{-5}$, $\mu = 1.5$, $m = 800$)

| IT | # pairs | Max of Θ | Elapsed time(s) |
|---|---|---|---|
| 1 | 265 | 1.8E-01 | 210 |
| 2 | 402 | 3.4E-03 | 317 |
| 3 | 402 | 2.8E-04 | 403 |
| 4 | 402 | 2.7E-04 | 483 |
| 5 | 402 | 2.7E-04 | 569 |
| 6 | 402 | 2.7E-04 | 669 |

**Table 6** Exam-1 ($n = 10$, $g_{\mathrm{S}} = 1\mathrm{E}{-5}$, $\mu = 1.5$, $m = 800$)

| IT | # pairs | Max of Θ | Elapsed time(s) |
|---|---|---|---|
| 1 | 286 | 1.7E-01 | 220 |
| 2 | 402 | 2.1E-03 | 347 |
| 3 | 402 | 2.7E-04 | 457 |
| 4 | 402 | 2.6E-04 | 567 |
| 5 | 402 | 2.6E-04 | 677 |
| 6 | 402 | 2.6E-04 | 801 |

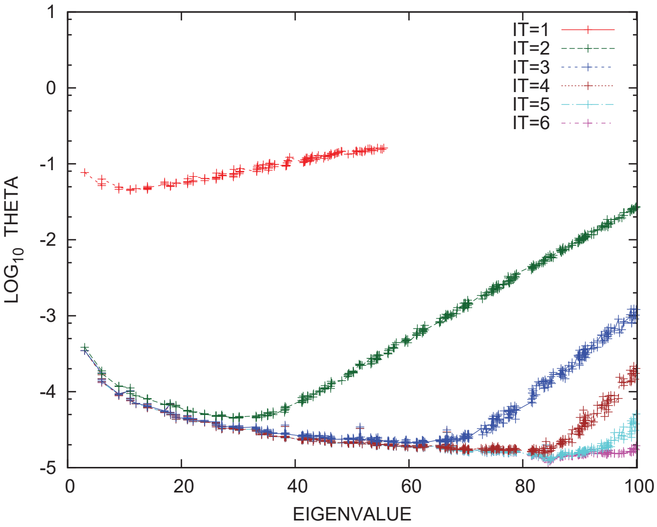

**Fig. 3** Exam-1 : Eigenvalue vs. relative residual ( $n = 4$, $g_S = 1\mathrm{E}{-5}$, $\mu = 1.5$, $m = 800$ ).

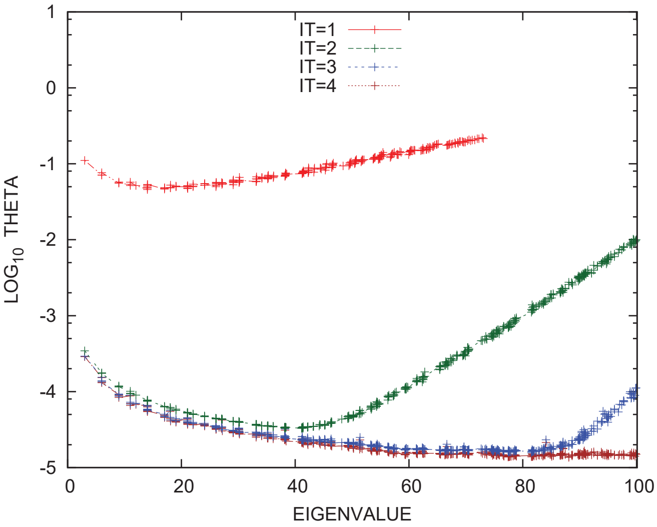

**Fig. 4** Exam-1 : Eigenvalue vs. relative residual ( $n = 6$, $g_S = 1\mathrm{E}{-5}$, $\mu = 1.5$, $m = 800$ ).

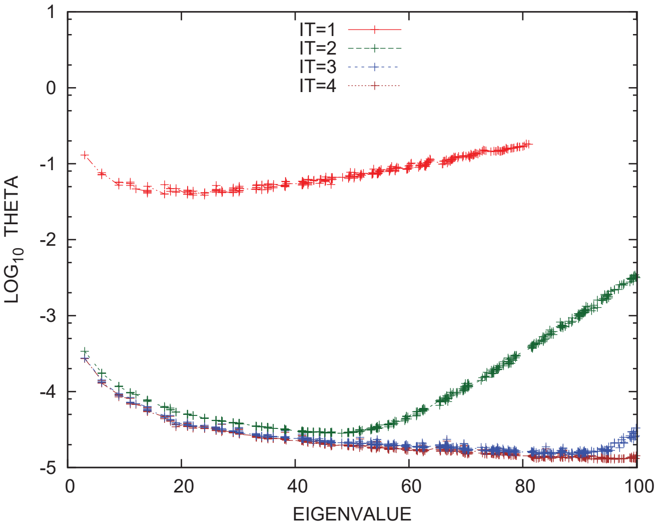

**Fig. 5** Exam-1 : Eigenvalue vs. relative residual ( $n = 8$, $g_S = 1\mathrm{E}{-5}$, $\mu = 1.5$, $m = 800$ ).

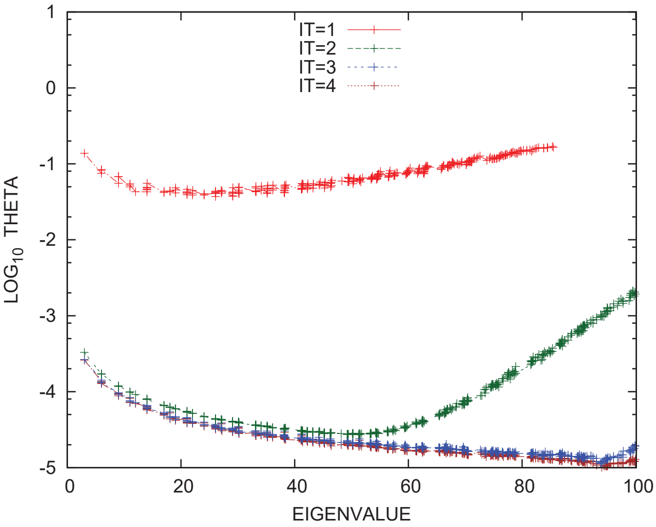

**Fig. 6** Exam-1 : Eigenvalue vs. relative residual ( $n = 10$, $g_S = 1\mathrm{E}{-5}$, $\mu = 1.5$, $m = 800$ ).

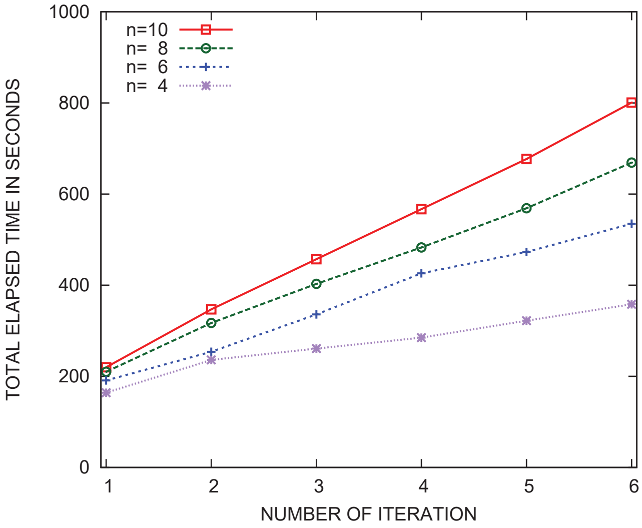

**Fig. 7** Exam-1 : Number of iteration vs. total elapsed time in seconds (filters consist of a single resolvent with a real shift, and initially $m = 800$ vectors are filtered ).

### 5.2 Exam-2: Approximate Eigenpairs with Internal Eigenvalues

In this example Exam-2, we try to solve those eigenpairs whose eigenvalues are in the interval $[a, b] = [100, 200]$. There are 801 such eigenpairs. Since their eigenvalues are interior of the eigenvalue distribution, we used an imaginary shift for the resolvent of each filter for this example.

There are 1,192 eigenvalues in the interval $[a', b'] = [75, 225]$, which is the union of the pass-band and transition-bands of the filter with $\mu = 1.5$. Since it is desirable to use more initial vectors than that number, we set the number of initial vectors to $m = 1,300$ for calculations.

With $g_S = 10^{-5}$ and the degree of the filter $n$ is increased from 4 to 10 by 2 (four tables from Table 7 to Table 10, and four figures from Figure 8 to Figure 11), the number of eigenpairs obtained is wrong and different from the correct number 801 for the case of IT = 1, but numbers are all correct for cases IT is 2 or more. And if we look into graphs in four figures which correspond to degrees $n$ from 4 to 10 by 2, we can recognize that for each case the relative residual of eigenpairs decreased well as the number of iteration IT increased (Since in each of these four figures, the graph of IT = 3(blue) and the graph of IT = 4(brown) overlap almost, therefore at IT = 3 the refinement has been finished already).

In each of tables (from Table 7 to Table 10), we also show the total elapsed time consumed to obtain approximate eigenpairs by IT times applications of the filter including $B$-reorthonormalizations to a set of $m$ vectors and also including the complex Cholesky decomposition once to prepare the filter. The graph is also shown in Figure 12.

**Table 7** Exam-2 ($n$ = 4, $g_S$ = 1E−5, $\mu$ = 1.5, $m$ = 1300)

| IT | # pairs | Max of Θ | Elapsed time(s) |
|---|---|---|---|
| 1 | 700 | 3.1E-01 | 413 |
| 2 | 801 | 2.3E-03 | 672 |
| 3 | 801 | 3.6E-05 | 790 |
| 4 | 801 | 2.2E-05 | 964 |

**Table 8** Exam-2 ($n$ = 6, $g_S$ = 1E−5, $\mu$ = 1.5, $m$ = 1300)

| IT | # pairs | Max of Θ | Elapsed time(s) |
|---|---|---|---|
| 1 | 800 | 3.3E-01 | 479 |
| 2 | 801 | 2.2E-04 | 782 |
| 3 | 801 | 2.3E-05 | 991 |
| 4 | 801 | 2.3E-05 | 1,189 |

**Table 9** Exam-2 ($n$ = 8, $g_S$ = 1E−5, $\mu$ = 1.5, $m$ = 1300)

| IT | # pairs | Max of Θ | Elapsed time(s) |
|---|---|---|---|
| 1 | 824 | 3.3E-01 | 564 |
| 2 | 801 | 8.1E-05 | 920 |
| 3 | 801 | 3.4E-05 | 1,157 |
| 4 | 801 | 3.3E-05 | 1,417 |

**Table 10** Exam-2 ($n$ = 10, $g_S$ = 1E−5, $\mu$ = 1.5, $m$ = 1300)

| IT | # pairs | Max of Θ | Elapsed time(s) |
|---|---|---|---|
| 1 | 828 | 2.7E-01 | 629 |
| 2 | 801 | 5.9E-05 | 1,028 |
| 3 | 801 | 3.6E-05 | 1,342 |
| 4 | 801 | 3.6E-05 | 1,657 |

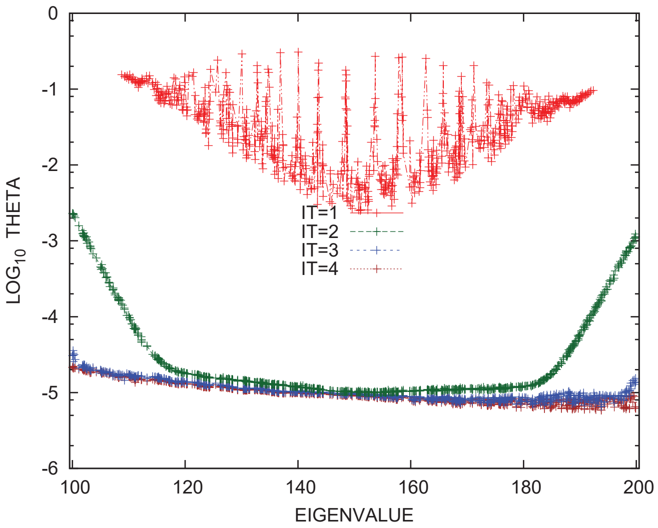

**Fig. 8** Exam-2 : Eigenvalue vs. relative residual ( $n$ = 4, $g_S$ = 1E−5, $\mu$ = 1.5, $m$ = 1300 ).

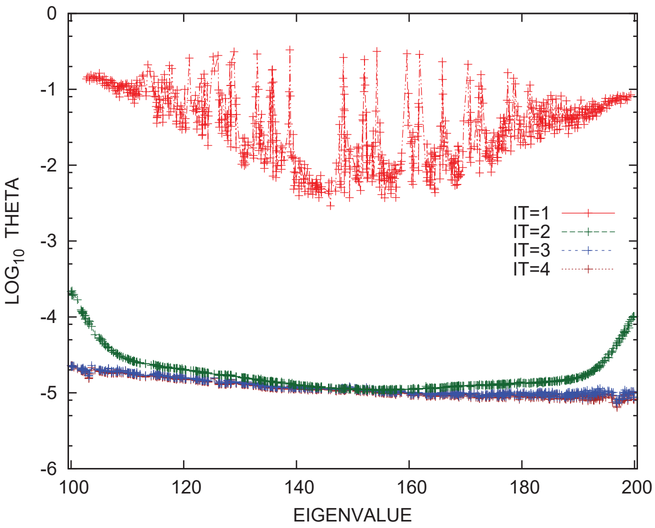

**Fig. 9** Exam-2 : Eigenvalue vs. relative residual ( $n$ = 6, $g_S$ = 1E−5, $\mu$ = 1.5, $m$ = 1300 ).

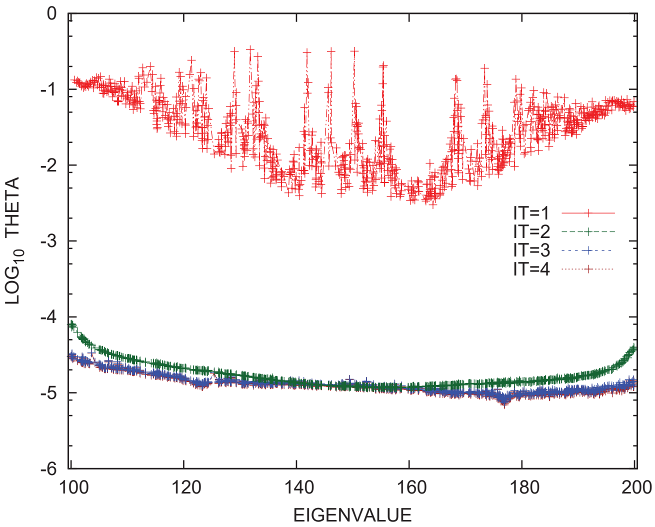

**Fig. 10** Exam-2 : Eigenvalue vs. relative residual ( $n$ = 8, $g_S$ = 1E−5, $\mu$ = 1.5, $m$ = 1300 ).

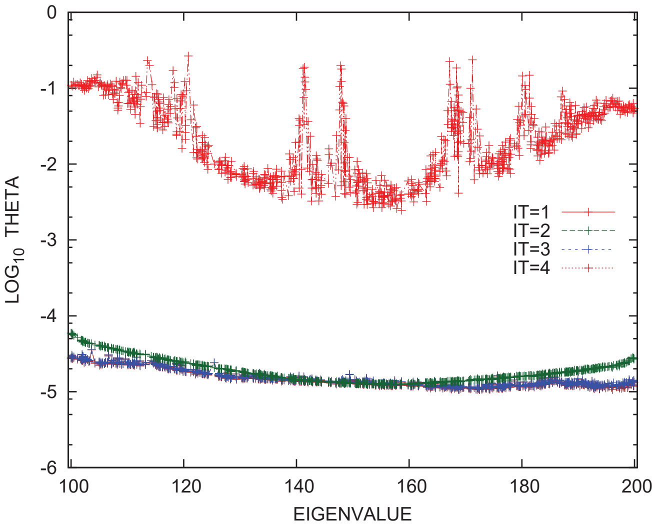

**Fig. 11** Exam-2 : Eigenvalue vs. relative residual ( $n$ = 10, $g_S$ = 1E−5, $\mu$ = 1.5, $m$ = 1300 ).

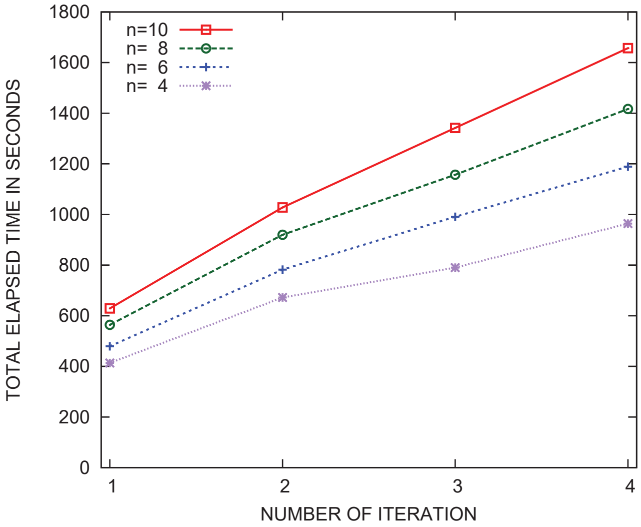

**Fig. 12** Exam-2 : Number of iteration vs. total elapsed time seconds (filters consist of a single resolvent with an imaginary shift, and initially $m = 1,300$ vectors are filtered ).

## 6. Conclusion

We made some experiments to solve those eigenpairs of a real symmetric-definite generalized eigenproblem whose eigenvalues are in the specified interval by using a filter.

Filters we used are composed of an action of a single resolvent. For the shift of resolvent, we can use a real number to solve those eigenpairs whose eigenvalues are at the lower-end of the eigenvalue distribution. When we use an imaginary number for the shift, the interval for eigenvalues to be solved may be placed anywhere. The filter is a real polynomial of a single resolvent whose shift is real, or it is a real polynomial of an imaginary part of a single resolvent whose shift is imaginary. When the degree of the real polynomial is $n$, there are $n$ applications of the resolvent inside an application of the filter. We used a Chebyshev polynomial to represent the real polynomial, which made the filter design simple, and also an application of the filter can be calculated by using the three-term recursion relation.

An application of a resolvent to a vector is calculated by solving a system of linear equations whose coefficient is the shifted matrix made from both matrices of the given generalized eigenproblem. We assumed that the system of linear equations is solved by some direct method by using decomposition of the coefficient matrix. Since our filter consists of a single resolvent, we need to make a matrix decomposition only once, and matrix-factors are hold and used sequentially $n$ times inside the filtering to solve a set of system of linear equations with a common matrix coefficient but $m$ different right-hand-sides, here $n$ is degree of the polynomial of the filter and $m$ is the number of vectors to be filtered.

By the use of a single resolvent for the filter instead of many, we reduced both costs to compute matrix decompositions and especially to store matrix-factors. However, those filters composed of a single resolvent are not good in uniformity of transfer-rate in the pass-band if they are compared to filters composed of many resolvents, especially when the precision of numbers and arithmetic operations used in computation is low.

The set of initial vectors, which is generated from random numbers, is $B$-orthonormalized and then filtered to give another set of vectors, to which we analyze and try to extract approximate eigenvectors. If uniformity of transfer-rate of the filter in the pass-band is not good, approximate eigenpairs which are required may be inaccurate or some of them may be lost, especially when the precision used in computation is low.

However, in the similar way as “orthogonal iteration” [4], [12], [13] which is a well-known method for many years, initially a set of vectors is generated from random numbers, and then the combination of orthonormalization and filtering is applied a few times to the set. The orthonormalization prevents the tendency of those vectors to become linearly dependent, and the filtering reduces well those eigenvectors which are not required. By this refinement, the set of vectors becomes a better approximate basis of the invariant-subspace which is spanned by required eigenvectors. From the set of refined vectors, we construct a basis of approximate invariant-subspace whose condition is good. Then, the Rayleigh-Ritz procedure is applied to the basis to obtain approximate eigenpairs required as Ritz pairs.

We made some experiments for a real symmetric-definite generalized eigenproblem of banded system whose matrix size is $210,000$ and lower-bandwidth is $3,051$, which comes from a FEM discretization of the Laplacian eigenproblem in a cube with zero-Dirichlet boundary condition at the surface of the cube. From results of experiments which used only single-precision for computations, even we used a filter whose characteristics were not so good because it was composed of only a single resolvent in order to reduce requirements for computer resources, we found present approach of iterative refinement worked very well to solve eigenpairs required.

### References

[1] Austin, A. P. and Trefethen, L. N.: Computing eigenvalues of real symmetric matrices with rational filters in real arithmetic, *SIAM J. Sci. Comput*, Vol. 37, No. 3, pp. A1365–A1387 (2015).

[2] Demmel, J. and Veselic, K.: Jacobi’s method is more accurate than QR, *SIAM J. Matrix Anal. Appl*, Vol. 13, pp. 1204–1245 (1992).

[3] Galgon, M., Krämer, L. and Lang, B.: The FEAST algorithm for large eigenvalue problems, *PAMM· Proc. Appl. Math. Mech.*, Vol. 11, No. 1, pp. 747–748 (2011).

[4] Golub, G. H. and van Loan, C. F.: *Matrix Computations*, The John Hopkins Univ. Press, 4 edition (2013).
(§8.2.4:‘Orthogonal Iteration’).

[5] Güttel, S., Polizzi, E., Tang, P. T. P. and Viaud, G.: Zolotarev quadrature rules and load balancing for the FEAST eigensolver, *SIAM J. Sci. Comput*, Vol. 37, No. 4, pp. A2100–A2122 (2015).

[6] Jed A. Duersch, Meiyue Shao, C. Y. and Gu, M.: A robust and efficient implementation of LOBPCG, *SIAM J. Sci. Comput.*, Vol. 40, No. 5, pp. C655–C676 (2018).

[7] Murakami, H.: Filter diagonalization method by using a polynomial of a resolvent as the filter for a real symmetric-definite generalized eigenproblem, *Eigenvalue Problems: Algorithms, Software and Applications in Petascale Computing* (Sakurai, T., Zhang, S., Imamura, T., Yamamoto, Y., Kuramashi, Y. and Hoshi, T., eds.), LNCSE, Vol. 117, Springer, pp. 368–394 (2018).

[8] Murakami, H.: Filters consists of a few resolvents to solve real symmetric-definite generalized eigenproblems, *Japan J. Indust. Appl. Math.*, Vol. 36, No. 2, pp. 579–618 (2019).

[9] Murakami, H.: Improvement of approximate pairs of a real symmetric-definite generalized eigenproblem by iterative applications of a filter *(written in Japanese)*, *IPSJ Transaction of Advanced Computing Systems (ACS65)*, Vol. 12, No. 3, pp. 14–33 (2019).
(written in Japanese).

[10] Polizzi, E.: A density matrix-based algorithm for solving eigenvalue problems, *Phys. Rev. B*, Vol. 79, No. 1, pp. 115112–115117 (2009).

[11] Rutishauser, H.: The Jacobi Method for Real Symmetric Matrices, Vol. 9, No. 1, pp. 1–10 (1966).

[12] Rutishauser, H.: Computational aspects of F. L. Bauer's simultaneous iteration method, *Numer. Math.*, Vol. 13, No. 1, pp. 4–13 (1969).

[13] Rutishauser, H.: Simultaneous iteration method for symmetric matrices, *Numer. Math.*, Vol. 16, No. 3, pp. 205–223 (1970).

[14] Stathopoulos, A. and Wu, K.: A block orthogonalization procedure with constant synchronization requirements, *SIAM J. Sci. Comput.*, Vol. 23, No. 6, pp. 2165–2182 (2002).

[15] Toledo, S. and Rabani, E.: Very large electronic structure calculations using an out-of-core filter-diagonalization method, *J. Comput. Phys.*, Vol. 180, No. 1, pp. 256–269 (2002).