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 (Tokyo Metropolitan University)

- By using a filter, we solve approximate eigenpairs of a real symmetric-definite GEVP whose eigenvalues are within a specified interval.
- We assume the system of linear equations, which gives the action of a resolvent, is solved by some direct method.
- The filter is a polynomial of a single resolvent in order to reduce both costs to factor the matrix and to store the factors.
  But, the transfer function of such a filter is not good in shape, and residuals of approximate eigenpairs will not be small.
- Vectors to span an approximate invariant subspace are improved by iterations of the combination of orthonormalization and filtering.

Filters composed of a single resolvent

- For a real symmetric-definite GEVP $Av = \lambda Bv$, we solve approximate eigenpairs whose eigenvalues are within the specified interval $[a, b]$ by using a filter.
- Our filter is composed of resolvents $\mathcal{R}(\rho_i) \equiv (A - \rho_i B)^{-1} B$ whose shift ρ_i are complex numbers.
  For a given vector x, an application of the resolvent $y \leftarrow \mathcal{R}(\rho)x$ reduces to solve $C(\rho)y = Bx$ for y whose coefficient is the shifted matrix $C(\rho) \equiv A - \rho B$, which in present study is solved by some direct method.
- The shifted matrix $C(\rho)$ is real-symmetric when ρ is real, and it is real-symmetric and positive-definite when ρ is real and less than the minimum eigenvalue.

- The matrix $C(\rho)$ is complex-symmetric non-singular when ρ is imaginary. To solve a symmetric system of linear equations, we make an LDL^T decomposition and forward and backward substitutions.
- Both amounts of *computation to factor matrices* and especially *storage to hold factors of matrices* tend to limit the calculation when a large size problem is solved under limited computing resources. Both are proportional to the number of resolvents used in the filter, thus it is desirable to reduce the number.
- We used two types of filters composed of only a single resolvent:
  1) $\mathcal{F} = g_s T_n(2\gamma \mathcal{R}(\rho) - I)$. 2) $\mathcal{F} = g_s T_n(2\gamma' \operatorname{Im} \mathcal{R}(\rho') - I)$.

- When the eigenvalue interval $[a, b]$ is located at the lower-end, the type-1 filter can be used and the shift ρ is real and less than the minimum eigenvalue. For the type-2 filter, the shift ρ' is imaginary and the interval can be placed anywhere. Here, g_s is the upper-bound of the transfer function magnitude in the stop-band, γ and γ' are real constants, and I denotes the identity operator.
- But, since only a single resolvent is used, transfer functions of these simple filters cannot have very good shapes.
  - Their transfer functions cannot have steep changes of values.
    $\Rightarrow$ μ, the ratio of the width of transition-bands to the width of the pass-band cannot be very small.

- Also, if g_s is set to a very small value, the max-min ratio of the transfer function within the pass-band $\lambda \in [a, b]$ will be larger.
- If this max-min ratio is very large, contained rates of required eigenvectors in the set of vectors tend to have different orders of magnitudes after the filtering.
- Thus, eigenvectors with larger transfer-rates dominate in a vector and eigenvectors with smaller transfer-rates reduce accuracy.
  $\Rightarrow$ Eigenvectors with smaller transfer-rates contained in filtered vectors tend to be inaccurate.
  $\Rightarrow$ Some approximate eigenpairs may fail to attain the level of required accuracy or lost.

Iterative Refinement of Eigenpairs by Using a Filter

- In the above, the filtering is assumed to make only once.
- Even the filter's transfer function is not good in shape, approximate eigenpairs can be improved if the combination of B-orthonormalization and filtering is iterated a small number (IT) of times by the following procedure:

1) Let $Y^{(0)}$ be an initial set of m random column vectors.
2) Iterate the followings for $i = 1, \ldots, \mathrm{IT}$
   $Y^{(i-1)}$ is B-orthonormalized to obtain $X^{(i)}$;
   $X^{(i)}$ is filtered to obtain $Y^{(i)}$.
3) Construct approximate eigenpairs from both sets $X^{(\mathrm{IT})}$ and $Y^{(\mathrm{IT})}$ considering the shape of the filter's transfer function.

- During the above iteration, m is updated to the effective rank of the set of vectors revealed by the B-orthonormalization in the step 2.
- The orthonormalization is introduced to prevent eigenvectors whose transfer-rates are relatively smaller from losing information by numerical rounding errors.
  The principle of using vector orthogonalization in each iteration step is well known and called as *orthogonal iteration*.

Test Problem for Experiments

- A 3-D Laplacian eigenproblem with zero-Dirichlet boundary for a cubic region whose length of a side is π :
$$-\Delta\,\Psi(x, y, z) = \lambda\,\Psi(x, y, z)\,. \qquad (1)$$
  By FEM discretization, a real symmetric-definite GEVP $A\,v = \lambda B\,v$ is obtained.
- Sides of the cube are equi-divided into $N_1 + 1$, $N_2 + 1$, $N_3 + 1$ sub-intervals to make finite elements (Fig. 1).

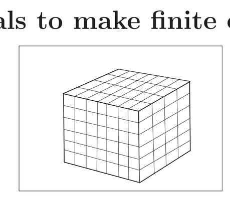


Figure 1: Concept of FE partitioning of a cube. Case $(N_1, N_2, N_3) = (3, 5, 6)$.

- Expansion basis inside an element : tri-linear functions.
- Matrix size of A and B : $N = N_1 N_2 N_3$ $(N_1 \le N_2 \le N_3)$.
  Lower band-width of A and B : $w_L = 1 + N_1 + N_1 N_2$.
- Filter diagonalization method is applied to solve approximate eigenpairs whose eigenvalues are within $[a, b]$.
- Eigenvalues of this test problem can be calculated exactly by simple formulae.
  Also the correct number of eigenvalues within any interval can be counted up.

Experiments of Iterative Refinements (in S-P)

FE partitionings of a cube : $(N_1, N_2, N_3) = (50, 60, 70)$.
A and B have size N=210,000 and lower-bandwidth w_L=3,051.

- Single-precision (IEEE 754 FP32, 7.2 digits precision) is used for numbers and arithmetics in calculations.
- Calculation in S-P has little margin for accuracy. However,
  - In recent years, attention has been paied to power saving through low-precision calculations.
  - There are systems that calculate much faster in S-P than in D-P.

Designs of Filters Used for Present Experiments

- For *lower-end* eigenpairs, the filter is a deg n Chebyshev polynomial of a resolvent with a *real* shift.
- For *interior eigenpairs*, the filter is a deg n Chebyshev poly of the *imaginary-part* of a resolvent with an *imaginary* shift.
- The filter's transfer function is specified by a set of parameters (n, μ, g_s), and here we always set $\mu = 1.5$.
- We prepared 6 designs of filters, both for lower-end eigenpairs and for interior ones.
  - Degree n is 4, and values of g_s are 1E-3, 1E-4 and 1E-5.
  - Value of g_s is 1E-5, and degree n are 6, 8 and 10.
- For 6 designs of filters of both types, values of g_p and g_s / g_p are shown (Tab. 1).

Table 1: Properties of 6 designed filters for lower-end eigenpairs and interior eigenpairs (μ=1.5) (g_s/g_p is the ratio of reduction per iteration.)

		for lower-end pairs		*for interior pairs*	
n	g_s	g_p	g_s / g_p	g_p	g_s / g_p
4	1E-3	1.9E-2	5.2E-2	7.2E-2	1.4E-2
4	1E-4	3.6E-3	2.8E-2	1.9E-2	5.3E-3
4	1E-5	5.3E-4	1.9E-2	3.7E-3	2.7E-3
6	1E-5	1.5E-3	6.5E-3	1.3E-2	8.0E-4
8	1E-5	2.6E-3	3.9E-3	2.1E-2	4.7E-4
10	1E-5	3.3E-3	3.0E-3	2.7E-2	3.7E-4

The larger the value of g_p, the better the filter.
The smaller the value of g_s/g_p, the better the filter.

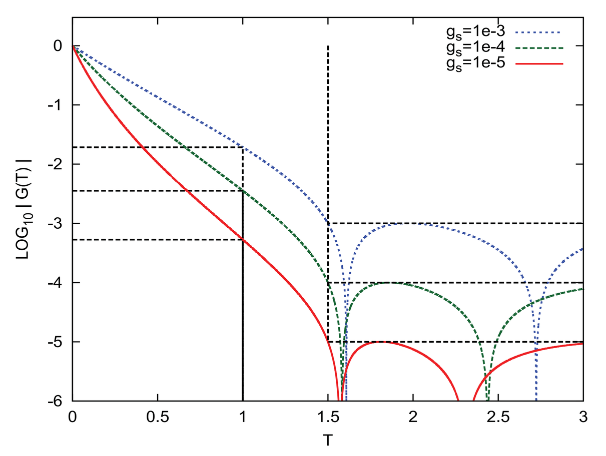


Figure 2: Log-plot of transfer function magnitude $|g(t)|$ $(n = 4)$ (Filter for *lower-end* eigenpairs)

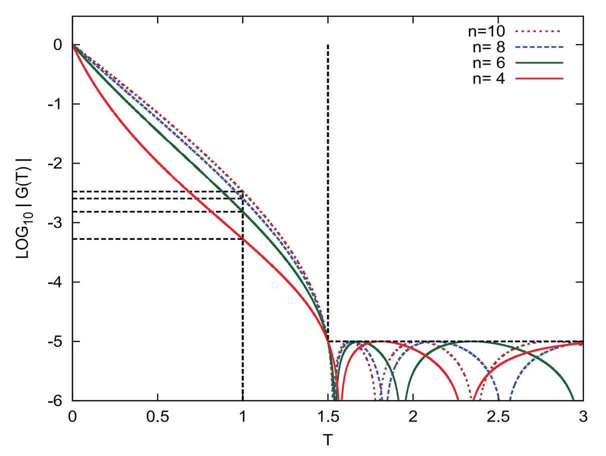


Figure 3: Log-plot of transfer function magnitude $|g(t)|$ (g_s = 1E-5) (Filter for *lower-end* eigenpairs)

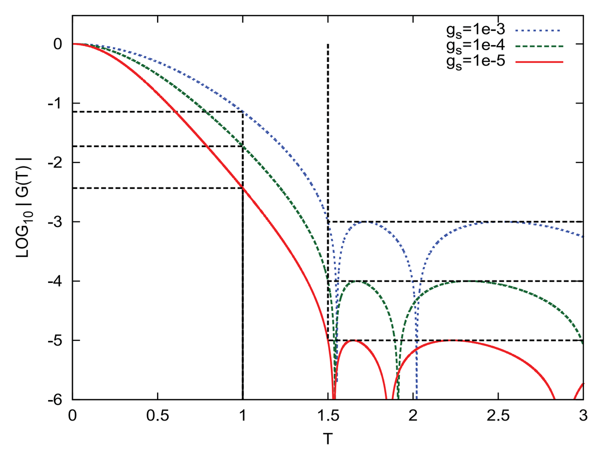


Figure 4: Log-plot of transfer function magnitude $|g(t)|$ ($n = 4$, right-half) (Filter for *interior* eigenpairs)

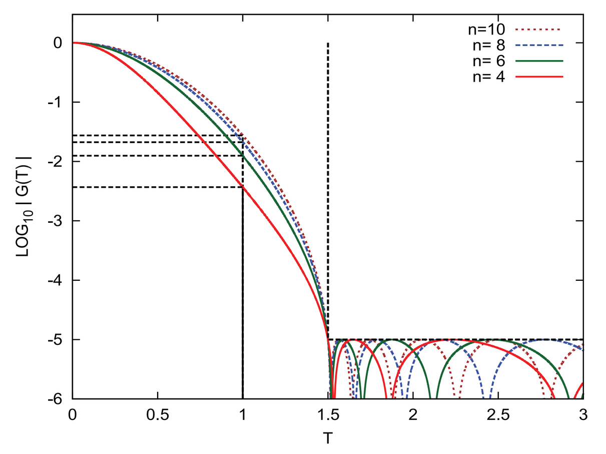


Figure 5: Log-plot of transfer function magnitude $|g(t)|$ (g_s = 1E-5, right-half) (Filter for *interior* eigenpairs)

Relative Residual of Eigenpair

- The quality of an approximate eigenpair (λ, v) is evaluated by the relative residual defined as :
$$\Theta \;\equiv\; \frac{\|\,A\,v - \lambda B\,v\,\|}{\|\,\lambda B\,v\,\|}\,. \qquad (2)$$
  Here, $\|\cdot\|$ is the 2-norm of a vector.
- The approximate eigenpair is accurate when Θ is small.
  - Θ does not depend on the vector normalization of v.
  - Θ is unchanged if both A and B are scaled by a factor.
- When ϕ is the angle between two vectors Av and λBv :
$$\sin\phi \;\le\; \Theta\,. \qquad (3)$$

7.2 (Ex-1): Solution of *Lower-end* Eigenpairs

- Within the lower-end interval $[a, b] = [0, 100]$, there are 402 eigenvalues to be solved.
- Within the union of the pass-band and the transition-band $[a, b'] = [0, 150]$, there are 764 eigenvalues.
- The number of vectors to be filtered is set to $m = 800$, which is more than 764 and to be sufficient.
- Results of experiments are shown (Tab. 2, Fig. 6).

Table 2: (Ex-1): num of iterations vs. num of approx eigenpairs and the max relative residuals. (The correct number of eigenpairs is 402.)

$n = 4$, $g_s =$ 1E-3

IT	# pairs	$\Theta_{\max}$
1	(5)	2.3E+00
2	(394)	2.3E-01
3	402	1.6E-02
4	402	9.6E-04
5	402	3.8E-04
6	402	3.8E-04

$n = 4$, $g_s =$ 1E-4

IT	# pairs	$\Theta_{\max}$
1	(82)	5.0E-01
2	402	6.8E-02
3	402	2.3E-03
4	402	4.1E-04
5	402	3.9E-04
6	402	4.1E-04

$n = 4$, $g_s =$ 1E-5

IT	# pairs	$\Theta_{\max}$
1	(139)	1.6E-01
2	402	2.7E-02
3	402	2.7E-03
4	402	8.7E-04
5	402	4.5E-04
6	402	4.5E-04

$n = 6$, $g_s =$ 1E-5

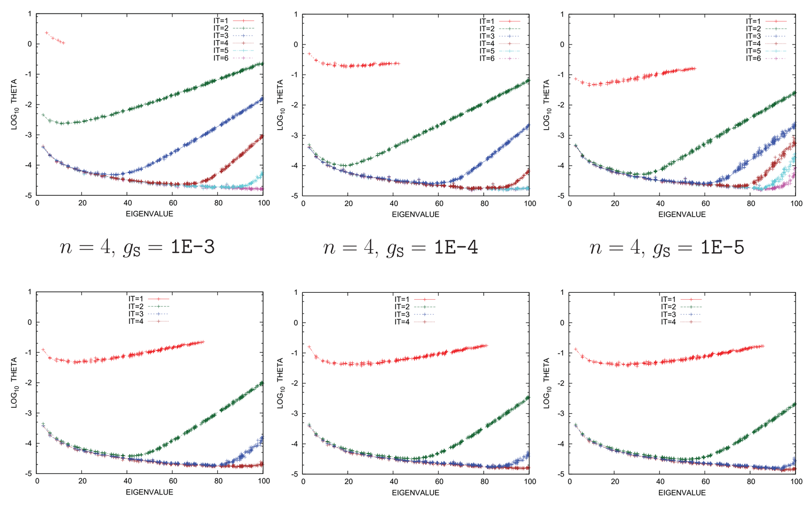
IT	# pairs	$\Theta_{\max}$
1	(222)	2.3E-01
2	402	1.1E-02
3	402	3.8E-04
4	402	3.8E-04

$n = 8$, $g_s =$ 1E-5

IT	# pairs	$\Theta_{\max}$
1	(264)	1.8E-01
2	402	3.6E-03
3	402	4.0E-04
4	402	3.9E-04

$n = 10$, $g_s =$ 1E-5

IT	# pairs	$\Theta_{\max}$
1	(287)	1.7E-01
2	402	2.2E-03
3	402	3.9E-04
4	402	4.0E-04



$n = 4$, g_s = 1E-3 $n = 4$, g_s = 1E-4 $n = 4$, g_s = 1E-5

$n = 6$, g_s = 1E-5 $n = 8$, g_s = 1E-5 $n = 10$, g_s = 1E-5

Figure 6: (Ex-1): eigenvalue vs. $\log_{10}$ of rel residual for each iteration (m=800 vectors are filtered. The correct number of eigenpairs is 402.)

7.3. (Ex-2): Solution of Interior Eigenpairs

- Within the interior interval $[a, b] = [100, 200]$, there are 801 eigenvalues to be solved.
- Within the union of the pass-band and transition-bands $[a', b'] = [75, 225]$, there are 1,192 eigenvalues.
- The number of vectors to be filtered is set to $m = 1,300$, which is more than 1,192 and to be sufficient.
- Results of experiments are shown (Tab. 3, Fig. 7).

Table 3: (Ex-2): num of iterations vs. num of approx eigenpairs and the max of relative residuals. (The correct number of eigenpairs is 801.)

$n = 4$, $g_s =$ 1E-3

IT	# pairs	$\Theta_{\max}$
1	(329)	1.9E-01
2	801	5.3E-02
3	801	9.4E-04
4	801	3.1E-05

$n = 4$, $g_s =$ 1E-4

IT	# pairs	$\Theta_{\max}$
1	(598)	1.8E-01
2	801	8.8E-03
3	801	5.6E-05
4	801	2.6E-05

$n = 4$, $g_s =$ 1E-5

IT	# pairs	$\Theta_{\max}$
1	(701)	3.0E-01
2	801	2.4E-03
3	801	2.9E-05
4	801	2.6E-05

$n = 6$, $g_s =$ 1E-5

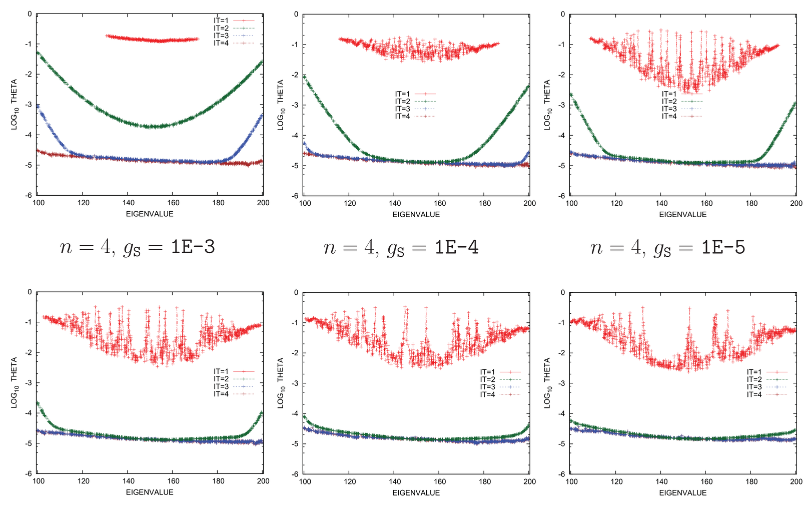
IT	# pairs	$\Theta_{\max}$
1	(800)	3.2E-01
2	801	2.2E-04
3	801	2.7E-05
4	801	2.7E-05

$n = 8$, $g_s =$ 1E-5

IT	# pairs	$\Theta_{\max}$
1	(825)	3.3E-01
2	801	8.3E-05
3	801	3.6E-05
4	801	3.5E-05

$n = 10$, $g_s =$ 1E-5

IT	# pairs	$\Theta_{\max}$
1	(828)	3.2E-01
2	801	6.0E-05
3	801	3.8E-05
4	801	3.8E-05



$n = 4$, g_s = 1E-3 $n = 4$, g_s = 1E-4 $n = 4$, g_s = 1E-5

$n = 6$, g_s = 1E-5 $n = 8$, g_s = 1E-5 $n = 10$, g_s = 1E-5

Figure 7: (Ex-2): eigenvalue vs. $\log_{10}$ of rel residual for each iteration (m=1300 vectors are filtered. The correct number of eigenpairs is 801.)

Conclusion

- We made some experiments for a banded real symmetric-definite generalized eigenproblem derived from FEM discretization of the Laplacian eigenproblem for a cubic region with zero-Dirichlet boundary condition.
- In experiments, we used filters composed of only a single resolvent in order to reduce computer resource requirements, but their transfer functions are not good in shapes.
- However, we found the present approach to improve eigenpairs iteratively worked well even when the calculation was made in single-precision.