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

Hiroshi Murakami (Tokyo Metropolitan University)

- We solve approximations of those pairs of real sym-def GEVP whose eigenvalues are in a specified interval by using a filter.
- The filter is constructed by resolvents. We assume some direct method is used to solve a large simultaneous linear equations which gives an action of a resolvent.
- In this study, the filter we used is a polynomial of a single resolvent so to reduce both costs to factor the matrix and to hold the factors, but properties of such kind of filter are not very good and residuals of approximate pairs will not be very small.
- Even filter's properties are not very good, the set of vectors spanning the approximate invariant-subspace is improved well by a few iterations of the combination of orthonormalization and filtering.

1. Introduction

- Those pairs of a real sym-def GEVP $Av = \lambda Bv$ whose eigenvalues are in the interval $[a, b]$ are solved approximately by using a filter.
- The filter consists of resolvents $\mathcal{R}(\rho_j) \equiv (A - \rho_j B)^{-1} B$.
- Types of filters $\mathcal{F}$ by construction are:
  - the real part of a linear combination of resolvents whose shifts ρ_j are complex.
    $\mathcal{F} \equiv c_\infty I + \mathrm{Re} \sum_{j=1}^{n} c_j \mathcal{R}(\rho_j)$.
  - a real polynomial P of a single resolvent whose shift ρ is real.
    $\mathcal{F} \equiv P(\mathcal{R}(\rho))$.
  - a real poly P of the imag-part of a single resolvent whose shift ρ' is imaginary.
    $\mathcal{F} \equiv P(\mathrm{Im}\,\mathcal{R}(\rho'))$.
  - mixed type (a real polynomial P of the real part of a linear combination of resolvents).
    $\mathcal{F} \equiv P\big(c_\infty I + \mathrm{Re} \sum_{j=1}^{k} c_j \mathcal{R}(\rho_j) \big)$.

- For a vector x, the action of a resolvent $y \leftarrow \mathcal{R}(\rho)\, x$ is to solve a simultaneous linear equations $C(\rho)\, y = B\, x$ for y, here $C(\rho) \equiv A - \rho B$ is the shifted matrix with a shift ρ. (The dominant part of the filtering calculation.)
- In this study, we assume : The simultaneous linear equations of a large size is solved by some direct method. $\Rightarrow$ Both amounts of computation to factor matrix and especially storage to hold the factors tend to be the bottle neck.
- Therefore, we desire to reduce the number of resolvents.

- Two kinds of filters which consist of a single resolvent:
  - a polynomial of a single resolvent whose shift is real.
  - a polynomial of the imag-part of a single resolvent whose shift is imaginary.

  But, transfer properties of these filters cannot be made very good even their shifts and polynomials are tuned.
  For example, it is not possible to make the steep discriminating property or the higher uniformity of transfer-rate magnitude in the pass-band.
  (The filter whose shift is imaginary has better properties compared with the one whose shift is real.)

- If we use 2 to 4 resolvents, properties of the filter can be improved very much. But, both amounts of computation to factor matrices and especially storage to hold their factors are proportional to the number of resolvents.
- Even properties of the filter is not very good, approximations could be improved by iteration.
  (If the same filter is used, matrix factors can be reused.)
- But, when the filtering is just repeated, magnitudes of eigenvectors are enhanced or depressed by powers of their transfer-rate magnitudes.
  $\Rightarrow$ Eigenvectors whose transfer-rate magnitudes are relatively small lose relative accuracy or may vanish.

- To prevent loss of accuracy, in every step of iteration we orthonormalize the set of vectors before it is filtered.
  This method is similar to *simultaneous inverse iteration* which has been used in the structure analysis.
  The principle is (*simultaneous*) *orthogonal iterations* [1][2][3].
- Therefore, we make experiments to improve accuracy of approximate pairs by repeating the combination of orthonormalization and filtering several times.
- To B-orthonormalize a set vectors, we use B-SVD, the singular value decomposition with metric B.

2. Simple Type Filters for Present Experiments

Filter is a polynomial of a single resolvent.
And a Chebyshev polynomial is used for the polynomial.
Transfer properties of this kind of filter is not very good.

- When interval $[a, b]$ for eigenvalue is lower-exterior, the shift ρ may be a real number, and the filter is (1).
$$\mathcal{F} \equiv g_s\, T_n\,(2\gamma\, \mathcal{R}(\rho) - I)\,. \tag{1}$$
- If an imaginary shift ρ' is used, the interval $[a, b]$ for eigenvalue may be placed anywhere, and the filter is (2).
$$\mathcal{F} \equiv g_s\, T_n\,\big(2\gamma'\, \mathrm{Im}\,\mathcal{R}(\rho') - I\big)\,. \tag{2}$$

$T_n(x)$: deg n Chebyshev poly. I: identity.
Im : operator to take imag-part. γ and γ': real consts.
g_s : max value of transfer-func magnitude in stop-band.

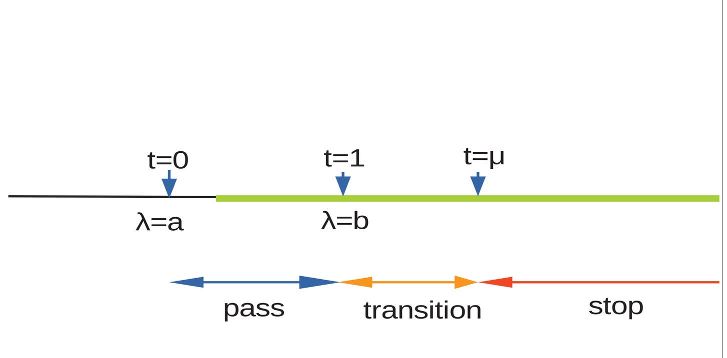


Figure 1: Eigenvalue λ and normalized coordinate t (for lower-exterior eigenvalues)

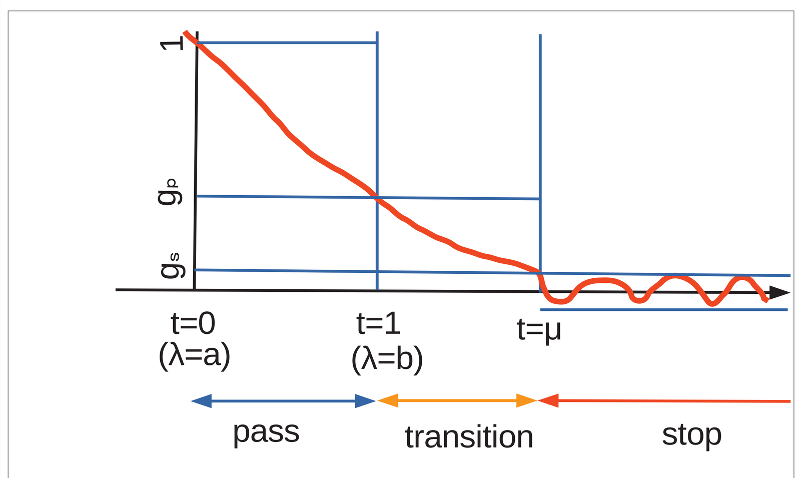


Figure 2: Conceptual shape of transfer-func magnitude $|g(t)|$ (for lower-exterior eigenvalues)

3. Designs of Present Filters

- The filter is specified by three parameters (n, μ, g_s).
  - n : degree of Chebyshev polynomial.
  - μ : normal coordinate of the edge of the stop-band.
  - g_s: bound of transfer-func magnitude in stop-band.
  - The range of transfer-func in the pass-band is $[g_p, 1]$.
- For each filtering, ratios of content rates of unrequired eigenvectors to content rates of required eigenvectors are reduced by factors of g_s / g_p or less.

3.1. Filter with a real shift

- When the interval $[a, b]$ for eigenvalue is lower-exterior, and $a \le \lambda_{\min}$.
- From the specified parameter set (n, μ, g_s), the shift ρ, the coefficient of resolvent γ and also g_p are calculated by (3).
$$\begin{cases} \sigma \leftarrow \mu / \sinh\left(\frac{1}{2n} \cosh^{-1} \frac{1}{g_s}\right), \\ \rho \leftarrow a - (b-a)\,\sigma, \\ \gamma \leftarrow (b-a)(\sigma + \mu), \\ g_p \leftarrow g_s \cosh\left\{2n \sinh^{-1} \sqrt{(\mu-1)/(1+\sigma)}\,\right\}. \end{cases} \tag{3}$$
- The filter $\mathcal{F}$ is given by (4).
$$\mathcal{F} \equiv g_s\, T_n\,(2\gamma\, \mathcal{R}(\rho) - I)\,. \tag{4}$$

3.2. Filter with an imaginary shift

- When the shift is imaginary, the interval $[a, b]$ for eigenvalue can be placed anywhere.
- From the specified parameter set (n, μ, g_s), the shift ρ', the coefficient of the resolvent γ' and also g_p are calculated by (5).
$$\begin{cases} \sigma \leftarrow \mu / \sinh\left(\frac{1}{2n} \cosh^{-1} \frac{1}{g_s}\right), \\ \rho' \leftarrow \frac{a+b}{2} + \left(\frac{b-a}{2}\right)\sigma\sqrt{-1}, \\ \gamma' \leftarrow \left(\frac{b-a}{2}\right)\frac{\mu^2+\sigma^2}{\sigma}, \\ g_p \leftarrow g_s \cosh\left\{2n \sinh^{-1} \sqrt{(\mu^2-1)/(1+\sigma^2)}\,\right\}. \end{cases} \tag{5}$$
- The filter $\mathcal{F}$ is given by (6).
$$\mathcal{F} \equiv g_s\, T_n\big(2\gamma'\, \mathrm{Im}\,\mathcal{R}(\rho') - I\big)\,. \tag{6}$$

4. Refinements of Vectors by Using a Filter

- Starting from a set of random vectors, we iteratively improve the set of vectors spanning the approximate invariant subspace.
- In each iteration the same filter is used, and the set of vectors is improved by the combination of B-orthonormalization and filtering.
- For B-orthonormalization, B-SVD with threshold is used.
  - We use 100 times the machine epsilon as the absolute value of the threshold.
  - Those singular vectors whose singular values are below the threshold are cut (removed).

4.1. Iterative refinement by using a filter

1. Factor shift matrix $C(\rho) = A - \rho B$ to prepare the filter $\mathcal{F}$.
2. $Y^{(0)} \Leftarrow$ a set of m random vectors.
3. for $i = 1, 2, \ldots, \mathtt{IT}$ do
   $X^{(i)} \Leftarrow B$-orthonormalization of $Y^{(i-1)}$; $Y^{(i)} \Leftarrow \mathcal{F}\, X^{(i)}$;
   enddo
   Note, if the effective rank is found reduced in orthonormalization, the number of vectors m in $X^{(i)}$ and $Y^{(i)}$ is updated.
4. We construct Z the basis of approximate invariant-subspace required from linear combinations of columns of $Y^{(\mathtt{IT})}$.
5. Approximate pairs of the original GEVP are the Ritz pairs obtained from Rayleigh-Ritz procedure applied to the basis Z.

5. GEVP for Experiments

- An EVP of 3-D Laplacian (7) is considered for a cube with a side length π with zero-Dirichlet boundary.
$$-\Delta\, \Psi(x, y, z) = \lambda\, \Psi(x, y, z)\,. \tag{7}$$
  By FEM discretization, we obtain a real symmetric-definite GEVP $A\, v = \lambda B\, v$.
- Sides of the cube are equi-divided into $N_1 + 1$, $N_2 + 1$, $N_3 + 1$ sub-intervals, to obtain finite elements (Fig. 3).

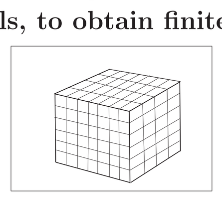


Figure 3: Concept of FE partitioning of a cube. Case $(N_1, N_2, N_3) = (3, 5, 6)$.

- Expansion basis inside an element: tri-linear functions in each direction.
- Size of both A and B : $N = N_1 N_2 N_3$ $(N_1 \le N_2 \le N_3)$.
  Lower bandwidth of A and B : $w_L = 1 + N_1 + N_1 N_2$.
- To this GEVP, the filter diagonalization method is applied to solve pairs with $\lambda \in [a, b]$.
- The exact eigenvalues of this GEVP are given by simple formulas.
- The number of exact eigenvalues in any interval can be obtained by counting.

6. Relative Residual of Eigenpair

- We use relative residual Θ given in (8) to evaluate the quality of the approximate pair (λ, v).
$$\Theta \equiv \frac{||\, A\, v - \lambda B\, v\, ||}{||\, \lambda B\, v\, ||}\,. \tag{8}$$
  If this number is small, the approximation of the eigenpair is good.
  - This value does not depend on the normalization of the vector v. It does not depend on the simultaneous scalings of both matrices A and B also.
  - For the vector's norm $||\cdot||$, we used 2-norm.
- When ϕ is the angle between vectors Av and λBv, then
$$\sin\phi \le \Theta\,. \tag{9}$$

7. Experiments of Iterative Refinements (in S-P)

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

- Calculations were made using IEEE 754 Single Precision (FP32, precision 7.2 digits).
- Fortran code written for double precision (FP64) was just converted to single precision (FP32) as :

```
REAL(KIND=8)    → REAL(KIND=4)
COMPLEX(KIND=8) → COMPLEX(KIND=4)
```

- With lower precision, there is less margin for accuracy. But we tried, because in recent years, power saving by lower precision calculation has attracted attention.

7.1. Designs of filters used for experiments

- For lower-exterior pairs, the filter is a deg n Chebyshev polynomial of a single resolvent whose shift is real.
- For interior pairs, the filter is a deg n Chebyshev-poly of an imag-part of a single resolvent whose shift is imag.
- Properties of the filter is specified by a set of parameters (n, μ, g_s) and we set $\mu = 1.5$ fixed.
- For both types of filters for lower-exterior pairs and interior pairs, we prepared six filter designs.
  - Degree n is 4, and values of g_s are 10^{-3}, 10^{-4} and 10^{-5}.
  - Value of g_s is 10^{-5}, and degrees n are 6, 8 and 10.
- For both types of filters, the values g_p and g_s / g_p are shown (Tab. 1) for six filter designs.

Table 1: Properties of designed six filters for lower-exterior pairs and interior pairs ($\mu = 1.5$) (g_s/g_p is the reduction rate per iteration.)

		for lower-ext pairs		for interior pairs	
n	g_s	g_p	g_s / g_p	g_p	g_s / g_p
4	10^{-3}	1.93×10^{-2}	5.19×10^{-2}	7.16×10^{-2}	1.40×10^{-2}
4	10^{-4}	3.56×10^{-3}	2.81×10^{-2}	1.88×10^{-2}	5.33×10^{-3}
4	10^{-5}	5.33×10^{-4}	1.88×10^{-2}	3.69×10^{-3}	2.71×10^{-3}
6	10^{-5}	1.53×10^{-3}	6.54×10^{-3}	1.25×10^{-2}	8.01×10^{-4}
8	10^{-5}	2.55×10^{-3}	3.92×10^{-3}	2.11×10^{-2}	4.73×10^{-4}
10	10^{-5}	3.34×10^{-3}	2.99×10^{-3}	2.74×10^{-2}	3.65×10^{-4}

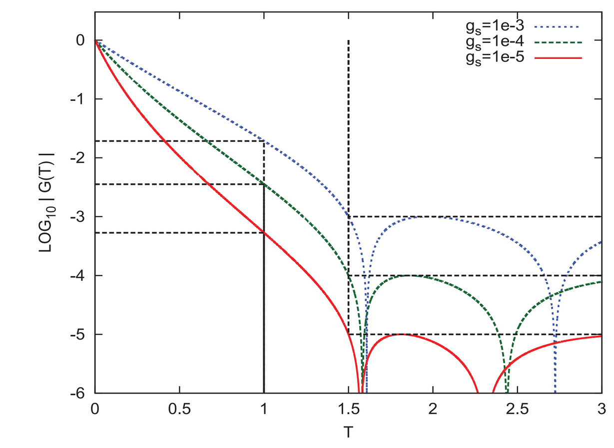


Figure 4: (for lower-exterior pairs) Transfer-func mag $|g(t)|$ ($n = 4$)

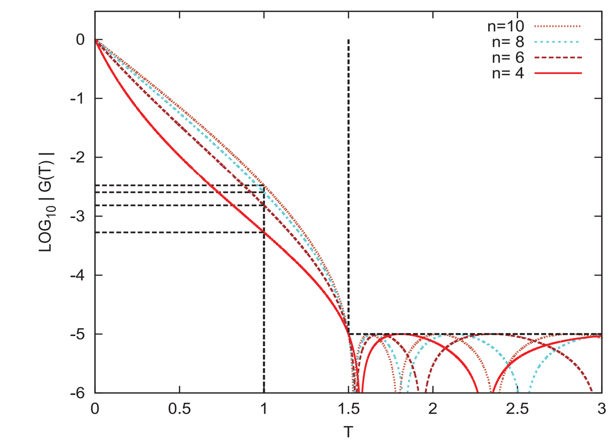


Figure 5: (for lower-exterior pairs) Transfer-func mag $|g(t)|$ ($g_s = 10^{-5}$)

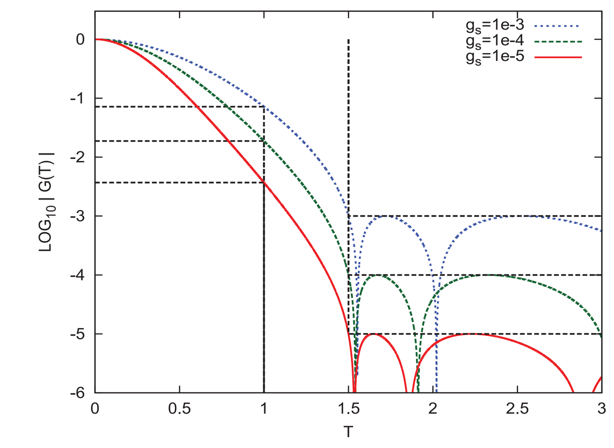


Figure 6: (for interior pairs) Transfer-func mag $|g(t)|$ ($n = 4$, right-half)

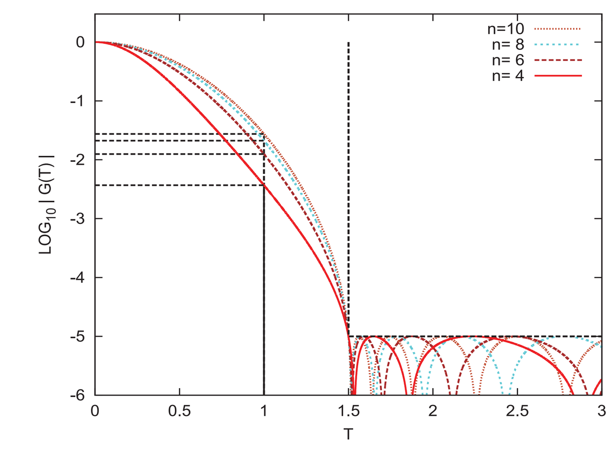


Figure 7: (for interior pairs) Transfer-func mag $|g(t)|$ ($g_s = 10^{-5}$, right-half)

7.2 (EX-1): Solution of lower-exterior eigenpairs

- We try to solve those 402 pairs whose eigenvalues are in the lower-exterior interval $[a, b] = [0, 100]$.
- There are 764 eigenvalues in $[a, b'] = [0, 150]$ which is the union of the pass-band and the transition-band.
- We use 800 for m the number of vectors for filtering, which is more than 764 and to be sufficient.
- Results of experiments are shown (Tab. 2, Fig. 8).

Table 2: EX-1: number of iterations vs. number of approximate pairs and the max of relative residuals (the number of true pairs is 402)

deg $n = 4$, $g_s = 10^{-3}$

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

deg $n = 4$, $g_s = 10^{-4}$

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

deg $n = 4$, $g_s = 10^{-5}$

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

deg $n = 6$, $g_s = 10^{-5}$

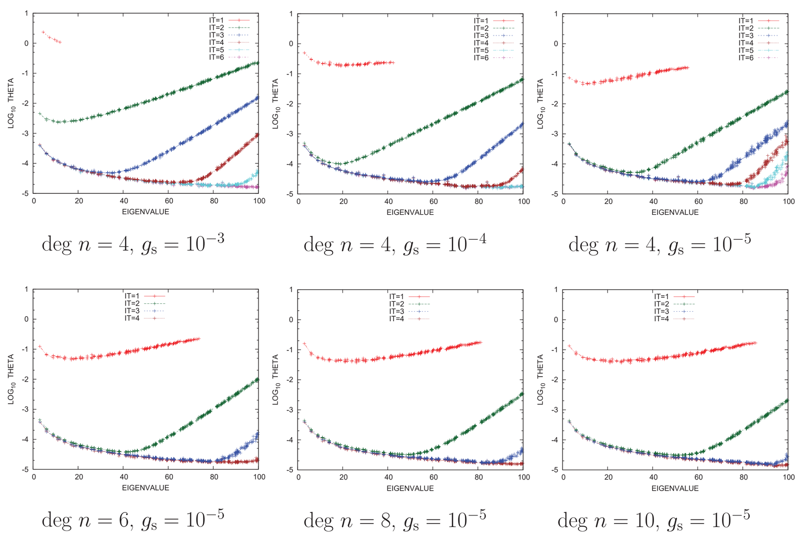
IT	# pairs	$\Theta_{\max}$
1	(222)	2.3E-01
2	402	1.1E-02
3	402	3.8E-04
4	402	3.8E-04

deg $n = 8$, $g_s = 10^{-5}$

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

deg $n = 10$, $g_s = 10^{-5}$

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



deg $n = 4$, $g_s = 10^{-3}$ deg $n = 4$, $g_s = 10^{-4}$ deg $n = 4$, $g_s = 10^{-5}$
deg $n = 6$, $g_s = 10^{-5}$ deg $n = 8$, $g_s = 10^{-5}$ deg $n = 10$, $g_s = 10^{-5}$

Figure 8: EX-1: eigenvalue vs. relative residual (the number of true pairs is 402. m=800 vectors are filtered.)

7.3. (EX-2): Solution of interior eigenpairs

- We try to solve those 801 pairs whose eigenvalues are in the interior interval $[a, b] = [100, 200]$.
- There are 1,192 eigenvalues in $[a', b'] = [75, 225]$ which is the union of the pass-band and transition-bands.
- We use 1,300 for m the number of vectors for filtering, which is more than 1,192 and to be sufficient.
- Results of experiments are shown (Tab. 3, Fig. 9).

Table 3: EX-2: number of iterations vs. number of approx pairs and the max of relative residuals (the number of true pairs is 801)

deg $n = 4$, $g_s = 10^{-3}$

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

deg $n = 4$, $g_s = 10^{-4}$

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

deg $n = 4$, $g_s = 10^{-5}$

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

deg $n = 6$, $g_s = 10^{-5}$

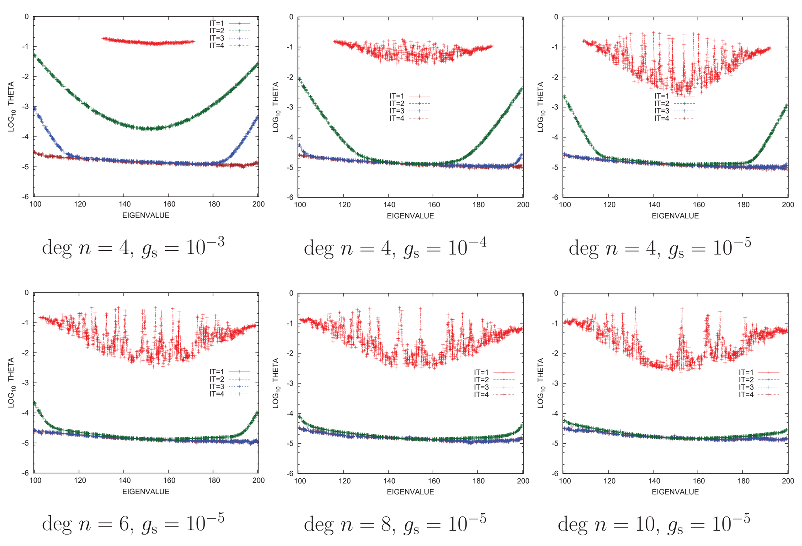
IT	# pairs	$\Theta_{\max}$
1	(800)	3.2E-01
2	801	2.2E-04
3	801	2.7E-05
4	801	2.7E-05

deg $n = 8$, $g_s = 10^{-5}$

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

deg $n = 10$, $g_s = 10^{-5}$

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



deg $n = 4$, $g_s = 10^{-3}$ deg $n = 4$, $g_s = 10^{-4}$ deg $n = 4$, $g_s = 10^{-5}$
deg $n = 6$, $g_s = 10^{-5}$ deg $n = 8$, $g_s = 10^{-5}$ deg $n = 10$, $g_s = 10^{-5}$

Figure 9: EX-2: eigenvalue vs. relative residual (the num of true pairs is 801. m=1,300 vectors are filtered.)

8. Conclusion

- By using a filter, we approximate pairs of a real-sym-def GEVP whose eigenvalues are in a specified interval.
- We used the filter which consists of a single resolvent to lessen both amounts of computation to factor matrices and especially storage to hold the matrix factors.
- A Chebyshev polynomial is used for the polynomial to make the filter design simple and also the filtering easy to implement by using the three-term recurrence.
- But a filter with such easy and simple construction cannot have a good transfer properties.
  $\Rightarrow$ Approximate pairs are not highly accurate or missed especially with low precision calculation.

- However, it has been found that even the filter's properties are not very good, repeating the combination of orthonormalization and filtering several times for a set of vectors improves the approximate pairs.
- Even in single precision calculations, approximate pairs of an eigenproblem are improved to have small relative residuals after several iterations.

REFERENCES

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

[2] Rutishauser, H.: Simultaneous iteration method for symmetric matrices, *Numer. Math.*, Vol.16, pp.205–223 (1970).

[3] Golub, G.H. and Van Loan, C.F.: *Matrix Computations*, 4th Ed., The John Hopkins Univ. Press (2013). (§8.2.4:'Orthogonal Iteration').