Inverse Iteration with Guaranteed Accuracy – a New Method for Computing Eigenvectors of Real Symmetric Tridiagonal Matrices

Anna Matsekh[1]

Russian Academy of Sciences, Novosibirsk (ICT)
Los Alamos National Laboratory (CCS-3)
matsekh@lanl.gov

Abstract. We discuss a new implementation of the inverse iteration method for computing eigenvectors of real symmetric tridiagonal matrices with guaranteed accuracy. The new method is an improved variant of the Godunov-inverse iteration method [Mat05], in which we use 'incomplete spectral deflation' [Mat04] in tight clusters, reducing the amount of reorthogonalization necessary when an orthogonal system of eigenvectors is required. We modify inverse iteration step to guarantee that computed eigenvector is a 'pseudoeigenvector' corresponding to the shift, chosen from the ϵ-pseudospectrum of the matrix. In many test cases C-implementation of this method exhibits performance, comparable and even superior to the LAPACK 3.0 [ABB$^+$99] implementations of the bisection and inverse iteration, divide and conquer, MRRR and QR methods for the symmetric tridiagonal eigenvalue problem.

1 Inverse Iteration with Guaranteed Accuracy

Suppose $\mathbb{R}^n$ is a real n-dimensional Euclidean subspace with fixed orthonormal basis, suppose $x \in \mathbb{R}^n$, A is a real tridiagonal $n \times n$ matrix, and $A = A^T$. Let $d_0, d_1, \dots, d_{n-1}$ represent elements of the main diagonal of A, and $b_0, b_1, \dots, b_{n-2}$ represent elements of the co-diagonals and $b_i \neq 0,\ i = 0, \dots, n-2$. Let $(x,\, x) = \sum_{k=0}^{n-1} |x_k|^2,\ \|x\| = \sqrt{(x,\, x)}$ and $\|A\| = \max_k |\lambda_k(A^T A)|$, where λ_k denotes k-th eigenvalue of a matrix, that is, $\|A\|$ denotes spectral norm of A. We require that approximate solution $(\tilde{x}_i, \tilde{\lambda}_i),\ 1 \leq i \leq n$ to the eigenvalue problem

$$A\, x_i = \lambda_i\, x_i$$

is such that $\max_i \|A\, \tilde{x}_i - \tilde{\lambda}_i\, \tilde{x}_i\| = O(\varepsilon_{\mathrm{mach}}\, \|A\|)$ and $\max_{i \neq j} |\tilde{x}_i^T\, \tilde{x}_j| = O(\varepsilon_{\mathrm{mach}})$, where $\varepsilon_{\mathrm{mach}}$ is the unit roundoff. Suppose we already found eigenintervals $[\alpha_i, \beta_i] \ni \lambda_i(A)$ and it is guaranteed that

$$\varepsilon_i \overset{\mathrm{def}}{=} |\beta_i - \alpha_i| \leq \max\{\varepsilon_{\mathrm{mach}},\, 3.0\, \|A\|_\infty\, \varepsilon_{\mathrm{mach}}\}.$$

We can then set $\tilde{\lambda}_i = (\alpha_i + \beta_i)/2.0$. In order to find a few $\tilde{\lambda}_i$ we can compute eigenintervals $[\alpha_i, \beta_i] \ni \lambda_i(A)$ using 'interval' implementation of the bisection method [GAKK93], [Mat05]. To find a large number of $\tilde{\lambda}_i$ we apply 'interval' bisection to the intervals $[\bar{\lambda}_i - \epsilon, \bar{\lambda}_i + \epsilon]$, where $\epsilon = 5.0 \, \|A\|_\infty \, \varepsilon_{\text{mach}}$ and $\bar{\lambda}_i$ is an eigenvalue approximation found applying LAPACK implementation of the root-free QR procedure (xSTERF). In order to determine approximate eigenvector $\tilde{x}_i$ we apply inverse iteration with the shift $\sigma_i \in [\alpha_i, \beta_i]$ and rescaling parameter $\tau_i > 0$ to the initial iterate $\tilde{x}_i^0 \in \mathbb{R}^n$ according to the following algorithm

Algorithm 1 (Inverse Iteration with Guaranteed Accuracy)

```
$z_i^0 = \tilde{x}_i^0 / \|\tilde{x}_i^0\|, \; k = 0$
**while** $\|z_i^k\| < \tau_i / \varepsilon_i$ **do**
    renormalize $\tilde{x}_i^k = \tau_i \, z_i^k / \|z_i^k\|$
    factor $A - \sigma_i \, I = L_i \, D_i \, L_i^T$
    solve $L_i \, D_i \, L_i^T \, z_i^{k+1} = \tilde{x}_i^k$
    $k = k + 1$
$\tilde{x}_i = z_i^k / \|z_i^k\|$.
```

In the algorithm 1 shift, rescaling parameter and termination criterion are chosen in accordance with the following proposition, based on the backward error analysis of the inverse iteration method [Ips97].

Proposition 1. *Suppose shift σ_i in the algorithm 1 is chosen such that $\|(A - \sigma_i \, I)^{-1}\| \geq 1/\varepsilon_i$, that is, σ_i belongs to the ϵ-pseudospectrum of A, and $\varepsilon_i \leq \tau_i \leq c \, \|A\|$. If $\|z_i^k\| \geq \tau_i / \varepsilon_i$, we can guarantee that $\|r_i^k\| = \|(A - \sigma_i \, I) \, \tilde{x}_i^k\| \leq \tau_i \, \varepsilon_i$, and $\|(A - \sigma_i \, I)^{-1}\| \geq \tau_i / \|r_i^k\| \geq 1/\varepsilon_i$, that is, we can guarantee that iterate z_i^k is a 'pseudoeigenvector', corresponding to the pseudoeigenvalue σ_i of the matrix A.*

Proof. Since $A - \sigma_i \, I$ is nonsingular and $\tilde{x}_i^k = (A - \sigma_i \, I)^{-1} \, (A - \sigma_i \, I) \, \tilde{x}_i^k = (A - \sigma_i \, I)^{-1} \, r_i^k$, we establish that $\|\tilde{x}_i^k\| \leq \|(A - \sigma_i \, I)^{-1}\| \, \|r_i^k\|$. But $\|\tilde{x}_i^k\| = \tau_i$, which means that

$$\|(A - \sigma_i \, I)^{-1}\| \geq \tau_i / \|r_i^k\|.$$

Noticing that $\|r_i^k\| = \|(A - \sigma_i \, I) \, x_i^k\| = \tau_i / \|z_i^k\| \, \|(A - \sigma_i \, I) \, z_i^k\| = \tau_i \, \|x_i^{k-1}\| / \|z_i^k\| = \tau_i^2 / \|z_i^k\|$, we establish that $\|z_i^k\| = \tau_i^2 / \|r_i^k\|$. This means that, as soon as $\|z_i^k\| \geq \tau_i / \varepsilon_i$, we can guarantee that residual r_i^k is small, that is, $\|r_i^k\| \leq \tau_i \, \varepsilon_i$, and

$$\|(A - \sigma_i \, I)^{-1}\| \geq \tau_i / \|r_i^k\| \geq 1/\varepsilon_i,$$

that is, we can guarantee that iterate z_i^k is a 'pseudoeigenvector', corresponding to the pseudoeigenvalue σ_i of the matrix A. $\triangle$

In our C-implementation of the algorithm 1 we set $\tau_i = \varepsilon_i$ in order to prevent overflow, while using $\sigma_i = \alpha_i$ as a shift. When computing a few eigenvectors we use choose $\sigma_i = \alpha_i$ if $|\alpha_i - \alpha_{i-1}| \geq \varepsilon_{\max}$, where $\varepsilon_{\max} = \max \varepsilon_i$, otherwise, in order to guarantee that $\sigma_i \in [\alpha_i, \, \beta_i]$, we set $\sigma_i = \min\left(\max\left(\beta_{i-1} + \varepsilon_i, \, \alpha_i\right), \beta_i\right)$. When an orthogonal system of approximate eigenvectors is required, we reorthogonalize $\tilde{x}_i$ against approximate eigenvectors $\tilde{x}_k, \;\; k < i$, already in the basis, applying Modified Gram-Schmidt reorthogonalization if $|\tilde{\lambda}_i - \tilde{\lambda}_k| \leq \gamma$, where $\gamma = \max |\lambda_i| \, \sqrt{\varepsilon_{\max}}$ if $\max \epsilon_i < \sqrt{\varepsilon_{\text{mach}}}$, and $\gamma = \max |\lambda_i| \, \sqrt[4]{\varepsilon_{\max}}$ otherwise.

If only one eigenvector, or an orthogonal system of eigenvectors, corresponding to well separated eigenvalues is required, we compute initial iterates $\tilde{x}_i^0$ using the same recursion as in the Godunov-inverse iteration algorithm [Mat05], that is,

$$\tilde{x}^0_{i,0} = 1, \quad \tilde{x}^0_{i,k} = -\mathrm{sign} b_{k-1} \frac{\tilde{x}^0_{i,k-1}}{P_{k-1}(\alpha_i, \beta_i)}, \quad k = 0, 1, \ldots, n-1,$$

where $P_0(\alpha_i, \beta_i), \ldots, P_{n-2}(\alpha_i, \beta_i) \overset{\text{def}}{=} P_0^+(\alpha_i), \ldots, P_{j-1}^+(\alpha_i), P_j^-(\beta_i), \ldots, P_{n-2}^-(\beta_i)$ is a two-sided Sturm sequence [GAKK93], [Mat05]. We use a variant of the 'incomplete spectral deflation' method [Mat04] to compute $\tilde{x}_i^0$ in tight clusters, as follows:

$$\begin{aligned}
\tilde{x}^0_{i,0} &= c_{-1}\, s_0\, s_1 \cdots s_{n-2-m}, \\
\tilde{x}^0_{i,k} &= c_{k-1}\, s_k\, s_{k+1} \cdots s_{n-2-m}, \quad k = 1, 2, \ldots, n-2-m, \\
\tilde{x}^0_{i,n-1-m} &= c_{n-2-m}, \\
\tilde{x}^0_{i,n-1-m+j} &\overset{\text{def}}{=} r_j, \quad 1 \le j \le m-1,
\end{aligned} \tag{1}$$

where $m = 0, 1, \ldots, n-2$ and r_j represent numbers from the random uniform distribution on $(0, 1)$. Parameters c_k, s_k are Givens rotation parameters, computed using two-sided Sturm sequences $P(\alpha_i, \beta_i)$ as follows [GAKK93]: $\mathrm{ctg}_k = -\mathrm{sign} b_k\, c_{k-1}/P_k(\alpha_i, \beta_i)$, $s_k = 1/\sqrt{1 + \mathrm{ctg}_k^2}$, $c_k = \mathrm{ctg}_k\, s_k$, $c_{-1} = 1$, $k = 0, 1, \ldots, n-2-m$. We can rewrite (1) in matrix form as follows: $\tilde{x}_i^0 = \tilde{y}_i^0 + u^{n-m}$, where $\tilde{y}_i^0 = C^m e^{n-m}$, $C^m = C_{n-2-m}\, C_{n-3-m} \cdots C_0$ is a chain of elementary Givens rotations $C_k = C_k(c_k, s_k)$, $e^{n-m} = (0, \ldots, 0, \underbrace{1}_{n-m}, 0, \ldots, 0)^T$ is a unit vector, and $u^{n-m} = (0, \ldots, 0, r_1, \ldots, r_{m-1})^T$. On each step m matrix A is replaced with the $n-1-m \times n-1-m$ tridiagonal matrix A^m such, that that partial vector $\bar{y}_i^0 = (\tilde{y}^0_{i,0}, \ldots, \tilde{y}^0_{i,n-1-m})$ is an approximate eigenvector of the $n-m \times n-m$ matrix [GAKK93]

$$\bar{A}^m = \begin{pmatrix} A^m & \\ & \tilde{\lambda}_i \end{pmatrix},$$

$$C^{mT} A^m C^m = C^{mT} C^{m-1T} \cdots C^{0T} A\, C^0 \cdots C^{m-1} C^m = \begin{pmatrix} A^m & & & \\ & \lambda_{i_0} & & \\ & & \ddots & \\ & & & \lambda_{i_p} \end{pmatrix},$$

where $\lambda_{n-1-m} \le \lambda_{i_0} \le \lambda_{i_p} \le \lambda_{n-1}$.

2 Example

Consider one-dimensional Poisson equation $-\partial^2 u/\partial x^2 = f \in [0, \pi]$ with Dirichlet boundary conditions $u(0) = 0$, $u(\pi) = \pi$. Finite difference approximation of this equation on a uniform mesh with step $h = \pi/N$ has symmetric tridiagonal $n \times n$ matrix T_h $(n = (N-1)^2)$, with main diagonal $1/h^2(2, 2, \ldots, 2)$ and co-diagonals $1/h^2(-1, -1, \ldots, -1)$. Analytical spectrum of T_h can be expressed explicitly as follows: $\lambda_k(T_h) = 2.0/h^2\,(1.0 - \cos[k\,\pi/(n+1)]),\ k = 1, 2, \ldots, n$. In the table below we present results of computing full spectral decomposition $T_h\,x_k = \lambda_k\,x_k,\quad k = 1, 2, \ldots, n$ for $h = \pi/96$ $(n = 9025)$, using LAPACK 3.0 routines dstedc (Divide and Conquer method), dstein (Inverse Iteration method) and dsteqr (QR method), development-LAPACK routine dstegr (MRRR method), and our C-implementation of the algorithm 1, which we call ginvit in the table below. The programs were tested on a four Intel Xeon 3.20 GHz CPU with total 4.0 GB of memory in Red Hat Linux 3.2.3 (gcc 3.2.3 compiler, default optimization). In the table $t_{\tilde{\lambda}}$ and $t_{\tilde{x}}$ represent processor time (in seconds) to compute respectively approximate eigenvalues and eigenvectors, $R(\tilde{\lambda}_k,\, \tilde{x}_k) = \max \|T_h\,\tilde{x}_k - \tilde{\lambda}_k\,\tilde{x}_k\|\,\|T_h\|^{-1}$, $O(\tilde{x}_k) = \max \|\tilde{x}_k^T\,\tilde{x}_k - 1\|$ and $\delta(\tilde{\lambda}_k) = \max |\tilde{\lambda}_k - \lambda_k|$.

	$R(\tilde{\lambda}_k,\, \tilde{x}_k)$	$O(\tilde{x}_k)$	$\delta(\tilde{\lambda}_k)$	$t_{\tilde{\lambda}}$	$t_{\tilde{x}}$	$t_{\tilde{\lambda}} + t_{\tilde{x}}$
ginvit	$2.05e-15$	$2.94e-14$	$2.04e-15$	9.54	34.57	44.11
dstegr	$3.77e-14$	$5.22e-12$	$5.22e-14$	3.75	51.51	55.26
dstedc	$2.61e-15$	$4.49e-14$	$2.39e-15$	4.23	1630.41	1634.64
dstein	$2.35e-15$	$1.37e-14$	$3.65e-16$	65.39	4816.92	4882.31
dsteqr	$9.44e-15$	$3.69e-14$	$5.72e-15$	5.74	5379.14	5384.88

Table 1. Full spectral decomposition of the matrix $T_h,\ h = \pi/96$ $(n = 9025)$.

References

ABB$^+$99. E. Anderson, Z. Bai, C. Bischof, S. Blackford, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney, D. Sorensen. *LAPACK Users' Guide*. SIAM, Philadelphia, third edition, 1999.

GAKK93. S. K. Godunov, A. G. Antonov, O. P. Kiriljuk, V. I. Kostin. *Guaranteed accuracy in numerical linear algebra*. Kluwer Academic Publishers Group, Dordrecht, 1993. ISBN 0-7923-2352-1. Translated and revised from the 1988 Russian original.

Ips97. Ilse C. F. Ipsen. Computing an eigenvector with inverse iteration. *SIAM Review*, 39(2):254–291, 1997.

Mat04. Anna Matsekh. Using spectral deflation to accelerate convergence of inverse iteration for symmetric tridiagonal eigenproblems. *Eighth Copper Mountain Conference on Iterative Methods, March 28-April 2*. Copper Mountain, CO, USA, 2004. URL `http://amath.colorado.edu/faculty/copper/2004/Abstracts/submission/matse099118.pdf`.

Mat05. Anna M. Matsekh. The Godunov-inverse iteration: A fast and accurate solution to the symmetric tridiagonal eigenvalue problem. *Applied Numerical Mathematics*, 54:208–221, 2005. URL `http://authors.elsevier.com/sd/article/S0168927404001977`.