On deflation for symmetric tridiagonal matrices

A.N. Malyshev*
Department of Informatics,
University of Bergen, 5020 Bergen, Norway

Abstract

K.V. Fernando developed an efficient approach for computation of an eigenvalue of a tridiagonal matrix corresponding to an approximate eigenvalue. We supplement Fernando's method with deflation procedures by Givens rotations. These deflations can be used in the Lanczos process and instead of the inverse iteration.

1 Introduction

Given a symmetric tridiagonal matrix

$$T = \begin{bmatrix} d_1 & b_1 & & & \\ b_1 & d_2 & b_2 & & \\ & \cdot & \cdot & \cdot & \\ & & b_{n-2} & d_{n-1} & b_{n-1} \\ & & & b_{n-1} & d_n \end{bmatrix}.$$

The identity matrix is as usual denoted by I.

Let λ be an approximate eigenvalue of T, i.e. $\det(T-\lambda I)$ is tiny. Assume additionally that $b_i \neq 0$, $i = 1, 2, \ldots, n-1$. Then it is very tempting to compute an eigenvector of T corresponding to λ by the following formulas:

$$x_1 = 1, \quad x_2 = -(d_1 - \lambda)/b_1,$$

$$x_i = -[b_{i-2}x_{i-2} + (d_{i-1} - \lambda)x_{i-1}]/b_{i-1}, \quad i = 3, \ldots, n.$$

However, this process is not always successful because the last equation $b_{n-1}x_{n-1} + (d_n - \lambda)x_n = 0$ may be strongly violated. J.H. Wilkinson tried to cure this problem in [9], [10] but could not succeed.

Only in summer 1983 S.K. Godunov found an elegant solution [4] to this problem. His approach is based on monotonicity of the Sturm sequences of symmetric tridiagonal matrices and uses a special "sewing" of two Sturm sequences: the one starting at the upper left corner of $T - \lambda I$ and the other starting at the lower right corner. A detailed exposition of Godunov's approach is found in [3].

A.N. Malyshev developed an efficient implementation of Godunov's algorithm [7] free of the annoying directional roundings for the floating-point numbers heavily used in [3].

In 1995 K.V. Fernando proposed a different approach to the problem under consideration and issued three technical reports summarized later in [2]. Fernando's solution is applicable in more

*This work was supported by the Research Council of Norway

general situations and looks slightly simpler than Godunov's solution. A survey of Fernando's approach is also found in [8].

The present paper was motivated by an attempt to combine Fernando's approach with the deflation techniques developed by S.K. Godunov and his collaborators. Among possible applications we saw Sorensen's implicitly restarted Lanczos [6].

In pursuit of this plan we discovered a new deflation procedure, which we refer to as "interior deflation". Recall that the common deflation occurs at one of the ends of the band of a tridiagonal matrix. In the interior deflation Givens rotations start from both ends of the band and meet somewhere inside it. The deflated eigenvalue emerges on the main diagonal of the transformed tridiagonal matrix at the meeting point. The rest of the transformed matrix forms a tridiagonal band with a bulge across the meeting point that can be chased in any direction. The interior deflation may be more attractive than the common one because it can be implemented without special tricks with extended exponents.

We note that our deflation algorithms are competitive with the inverse iteration [1]. They are robust and simple: no complications with multiple eigenvalues arise.

2 Polarizing a vector by chains of rotations

Let us denote by G_i an n-by-n Givens matrix

$$G_i = \begin{bmatrix} I & & & \\ & c_i & s_i & \\ & -s_i & c_i & \\ & & & I \end{bmatrix},$$

in which the unitary block $\begin{bmatrix} c_i & s_i \\ -s_i & c_i \end{bmatrix}$ stands at the intersection of rows i and $i+1$ and columns i and $i+1$. By a chain of rotations we call a matrix product

$$G = G_1 G_2 \dots G_{n-1}.$$

Let us say that a chain of rotations G polarizes an n-vector x if

$$G^* x = G_{n-1}^* \dots G_1^* x = \rho_n e_n = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ \rho_n \end{bmatrix}. \tag{1}$$

Here and henceforth $*$ denotes the Hermitian conjugation of matrices. It is easy to show that (1) implies the following representation:

$$G_{i-1}^* \dots G_1^* x = [\ 0 \ \ \dots \ \ 0 \ \ \rho_i \ \ x_{i+1} \ \ \dots \ \ x_n\ ]^T. \tag{2}$$

Using this representation we derive the system

$$\begin{bmatrix} c_i & s_i \\ -s_i & c_i \end{bmatrix}^* \begin{bmatrix} \rho_i \\ x_{i+1} \end{bmatrix} = \begin{bmatrix} 0 \\ \rho_{i+1} \end{bmatrix},$$

from which the equalities $\rho_i = s_i \rho_{i+1}$ and $x_{i+1} = c_i \rho_{i+1}$ follow. As a result,

$$x_i = c_{i-1}\rho_i = c_{i-1} s_i \rho_{i+1} = c_{i-1} s_i s_{i+1} \rho_{i+2} = \dots = c_{i-1} s_i \dots s_{n-1} \rho_n.$$

Thus (1) implies the formulas

$$\begin{aligned} x_1 &= \rho_n c_0 s_1 \dots s_{n-1}, \\ &\vdots \\ x_i &= \rho_n c_{i-1} s_i \dots s_{n-1}, \\ &\vdots \\ x_n &= \rho_n c_{n-1}, \end{aligned} \tag{3}$$

where $c_0 = 1$.

3 Deflation of Hessenberg matrices by chains of rotations

Recall that a matrix H is called lower Hessenberg if $H_{ij} = 0$ whenever $j > i + 1$.

Theorem 1 (O.P. Kirilyuk [5]) *Let H be an n-by-n lower Hessenberg matrix with an eigenvalue λ. Suppose that a corresponding eigenvector x satisfies the condition $|x_1|^2 + |x_2|^2 + |x_3|^2 \neq 0$. Let G be any chain of rotations polarizing the eigenvector x. Then the matrix $H_G = G^*HG$ is lower Hessenberg, $(H_G)_{n-1,n} = 0$, and $(H_G)_{n,n} = \lambda$.*

Remark 1 *The condition $|x_1|^2 + |x_2|^2 + |x_3|^2 \neq 0$ is satisfied, for example, when H is unreduced, i.e. $H_{i,i+1} \neq 0$ for all $i = 1, 2, \dots, n-1$.*

Proof of Theorem 1: Introduce the sequence of matrices

$$H_i = G_{i-1}^* \dots G_1^* H G_1 \dots G_i, \qquad i = 1, 2, \dots, n-1.$$

We will prove by induction that all H_i are lower Hessenberg. The matrix H_1 is trivially lower Hessenberg. If H_i is lower Hessenberg by the inductive assumption, then the matrix H_{i+1} is "almost" lower Hessenberg:

$$H_{i+1} = \begin{bmatrix} \times & \times & & & & & & \\ & \ddots & \ddots & & & & & \\ & & \times & \times & h_{i,i+2} & & & \\ & \ddots & & \times & \times & & & \\ & & & & \ddots & \ddots & & \\ \times & & & \ddots & & & \times & \times \\ \times & \times & & & & & & \times \end{bmatrix}.$$

Only the element $h_{i,i+2}$ in position $(i, i+2)$ can violate the structure.

From the identity $Hx = \lambda x$ we derive the equality

$$H_{i+1}(G_{i+1}^* \dots G_1^* x) = \lambda(G_i^* \dots G_1^* x).$$

With the help of representation (2) we write down the i-th row of this equality: $h_{i,i+2}\rho_{i+2} = 0$. Since $\rho_{i+2} \neq 0$ for all $i = 1, 2, \dots, n-2$, $h_{i,i+2} = 0$.

Finally, $H_G = G_{n-1}^* H_{n-1}$ is evidently lower Hessenberg and the remaining assertions of Theorem 1 follow from the identity $G^*HGe_n = \lambda e_n$. ∎

4 Deflation of symmetric tridiagonal matrices by chains of rotations

Let an n-by-n real symmetric tridiagonal matrix T_n have an eigenvalue λ and a corresponding real eigenvector x such that $\|(x_1\ x_2\ x_3)\| \neq 0$. Assume that a chain of rotations G with real parameters c_i, s_i polarizes the vector x. By Kirilyuk's theorem the matrix $T_G = G^T T_n G$ is lower Hessenberg and therefore tridiagonal by symmetry. Thus,

$$T_G = G^T T_n G = \begin{bmatrix} T_{n-1} & 0 \\ 0 & \lambda \end{bmatrix},$$

where T_{n-1} is real symmetric tridiagonal.

The above result is not very useful for the deflation computed with rounding errors. Assume that the rotation parameters c_i and s_i are computed from the components of x with small relative errors. Then the backward error analysis provides an identity $\widehat{T}_n \widehat{x} = \lambda \widehat{x}$, where $\widehat{T}_n$ is a tridiagonal perturbation of T_n, possibly unsymmetric, and $\widehat{x}$ is a perturbation of x. The computed rotation parameters satisfy (3) with $\widehat{x}$ instead of x. By Kirilyuk's theorem the matrix $H_G = G^T \widehat{T}_n G$ is lower Hessenberg, $(H_G)_{n,n} = \lambda$, and $(H_G)_{n-1,n} = 0$.

Let us split H_G into the sum

$$H_G = \begin{bmatrix} T_{n-1} & 0 \\ 0 & \lambda \end{bmatrix} + \Delta,$$

where T_{n-1} is symmetric tridiagonal and Δ is strictly lower triangular. Then the following identity holds:

$$G^T \widehat{T}_n G - G^T \widehat{T}_n^T G = \Delta - \Delta^T.$$

With the help of this identity we estimate $\|\Delta\|_2$:

$$\|\Delta\|_2 \leq \|\Delta\|_F = \frac{1}{\sqrt{2}} \|\Delta - \Delta^T\|_F \leq \sqrt{\frac{n}{2}} \|\Delta - \Delta^T\|_2 = \sqrt{\frac{n}{2}} \|\widehat{T}_n - \widehat{T}_n^T\|_2$$

$$\leq \sqrt{\frac{n}{2}} \left(\|\widehat{T}_n - T_n\| + \|T_n - \widehat{T}_n^T\|_2 \right) = \sqrt{2n} \|\widehat{T}_n - T_n\|_2.$$

As a result,

$$\|G^T T_n G - \begin{bmatrix} T_{n-1} & 0 \\ 0 & \lambda \end{bmatrix}\|_2 \leq \|G^T (T_n - \widehat{T}_n) G\|_2 + \|\Delta\|_2 \leq (1 + \sqrt{2n}) \|\widehat{T}_n - T_n\|_2. \qquad (4)$$

5 Computation of eigenvectors of tridiagonal vectors

Consider a tridiagonal matrix

$$T = \begin{bmatrix} d_1 & b_1 & & & \\ a_1 & d_2 & b_2 & & \\ & \cdot & \cdot & \cdot & \\ & & a_{n-2} & d_{n-1} & b_{n-1} \\ & & & a_{n-1} & d_n \end{bmatrix}. \qquad (5)$$

Let λ be an eigenvalue of T and assume that there exist the LDU and UDL factorizations of $T - \lambda I$:

$$T - \lambda I = L_+ D_+ U_+ = \begin{bmatrix} 1 & & & \\ l_1^+ & 1 & & \\ & \ddots & \ddots & \\ & & l_{n-1}^+ & 1 \end{bmatrix} \begin{bmatrix} d_1^+ & & & \\ & d_2^+ & & \\ & & \ddots & \\ & & & d_n^+ \end{bmatrix} \begin{bmatrix} 1 & u_1^+ & & \\ & 1 & \ddots & \\ & & \ddots & u_{n-1}^+ \\ & & & 1 \end{bmatrix}, \quad (6)$$

$$T - \lambda I = U_- D_- L_- = \begin{bmatrix} 1 & u_1^- & & \\ & 1 & \ddots & \\ & & \ddots & u_{n-1}^- \\ & & & 1 \end{bmatrix} \begin{bmatrix} d_1^- & & & \\ & d_2^- & & \\ & & \ddots & \\ & & & d_n^- \end{bmatrix} \begin{bmatrix} 1 & & & \\ l_1^- & 1 & & \\ & \ddots & \ddots & \\ & & l_{n-1}^- & 1 \end{bmatrix}. \quad (7)$$

We recall that factorization (6) is computed recursively as

$$\begin{aligned} &d_1^+ = d_1 - \lambda, \quad d_i^+ = d_i - \lambda - a_{i-1} u_{i-1}^+, \quad i = 2, \dots, n, \\ &u_i^+ = b_i / d_i^+, \quad i = 1, 2, \dots, n-1, \\ &l_i^+ = a_i / d_i^+, \quad i = 1, 2, \dots, n-1. \end{aligned} \quad (8)$$

Analogously, factorization (7) is computed by the formulas

$$\begin{aligned} &d_n^- = d_n - \lambda, \quad d_i^- = d_i - \lambda - b_i l_i^+, \quad i = n-1, \dots, 1, \\ &l_i^- = a_i / d_{i+1}^-, \quad i = n-1, \dots, 1, \\ &u_i^- = b_i / d_{i+1}^-, \quad i = n-1, \dots, 1. \end{aligned} \quad (9)$$

The existence of factorizations (6), (7) necessitates that

$$d_i^+ \neq 0 \quad \text{for } i = 1, 2, \dots, n-1, \quad \text{and } d_i^- \neq 0 \quad \text{for } i = n, \dots, 2. \quad (10)$$

The system of linear equations $(T - \lambda I)x = 0$ is singular but its solutions are easily found by means of factorizations (6) and (7). Having eliminated the k-th equation and putting $x_k = 1$ one can compute all other components of x. The vector $(x_1, \dots, x_k)^T$ satisfies the system

$$\begin{bmatrix} 1 & u_1^+ & & & \\ & 1 & u_2^+ & & \\ & & \ddots & \ddots & \\ & & & 1 & u_{k-1}^+ \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_k \end{bmatrix} = 0,$$

which implies that

$$x_k = 1, \qquad x_i = -u_i^+ x_{i+1}, \qquad i = 1, \dots, k-1. \quad (11)$$

Analogously, the vector $(x_k, \dots, x_n)^T$ satisfies the system

$$\begin{bmatrix} l_k^- & 1 & & \\ & \ddots & \ddots & \\ & & l_{n-1}^- & 1 \end{bmatrix} \begin{bmatrix} x_k \\ x_{k+1} \\ \vdots \\ x_n \end{bmatrix} = 0,$$

from which we derive that

$$x_k = 1, \qquad x_{i+1} = -l_i^- x_i, \qquad i = k, \dots, n-1. \quad (12)$$

Note that the k-th equation satisfies due to the singularity of the matrix $T - \lambda I$.

In reality, instead of a precise eigenvalue there is only an approximate one, and the computations are also carried out with rounding errors. Thus after computation of x by formulas (11), (12) instead of the identity $a_{k-1}x_{k-1} + (d_k - \lambda)x_k + b_k x_{k+1} = 0$ we get the residual

$$\begin{aligned} \gamma_k &= a_{k-1}x_{k-1} + (d_k - \lambda)x_k + b_k x_{k+1} \\ &= -a_{k-1}u_{k-1}^{+} + (d_k - \lambda) - b_k l_k^{-}. \end{aligned} \tag{13}$$

Equivalent formulas for γ_k are

$$\gamma_k = d_k^{+} + d_k^{-} - (d_k - \lambda) = d_k^{+} - b_k l_k^{-} = d_k^{-} - a_{k-1}u_{k-1}^{+}. \tag{14}$$

A remarkable fact discovered by Fernando is that $\min_i |\gamma_i|$ is very small when λ is a good approximate eigenvalue of T. Indeed, when $T - \lambda I$ is singular, $\min_i |\gamma_i| = \gamma_1 = \gamma_n = 0$. If $T - \lambda I$ is invertible, the conditions $(T - \lambda I)x = \gamma_k e_k$ and $x_k = 1$ imply that $1 = x_k = \gamma_k e_k^T (T - \lambda I)^{-1} e_k$ and $1/\gamma_k = e_k^T (T - \lambda I)^{-1} e_k$. We denote by e_k the k-th unit coordinate vector. Since $\sum_{i=1}^{n} \frac{1}{\gamma_i} =$ trace$(T - \lambda I)^{-1}$, $\frac{1}{\min_i |\gamma_i|} \geq \frac{1}{n} \left| \sum_{j=1}^{n} \frac{1}{\lambda_j - \lambda} \right|$. Therefore, if λ is close to one of λ_j, then $\min_i |\gamma_i|$ is tiny.

Let k be such that $|\gamma_k| = \min_i |\gamma_i|$. Then the vector x computed by (11), (12) will be an exact eigenvector of the perturbed matrix $T - \gamma_k e_k e_k^T$.

Further analysis of Fernando's method is found in [2] and [8].

6 Computation of the rotation parameters for the deflation of tridiagonal matrices

Consider an unreduced matrix T of the form (5). This means that $a_i \neq 0$, $b_i \neq 0$ for $i = 1, 2, \ldots, n-1$. Assume that a real number λ is an approximate eigenvalue of T and there exist the LDU and UDL factorizations of the matrix $T - \lambda I$. Let us determine k such that $|\gamma_k| = \min_i |\gamma_i|$ and compute an eigenvector x of the matrix $T - \gamma_k e_k e_k^T$ corresponding to λ.

The above conditions guarantee that $u_i^{+} \neq 0$ for $i = 1, 2, \ldots, k-1$ and

$$\frac{x_{i+1}}{x_i} = -\frac{1}{u_i^{+}}, \qquad i = 1, 2, \ldots, k-1. \tag{15}$$

Analogously, $l_i^{-} \neq 0$ for $i = k, \ldots, n-1$ and

$$\frac{x_{i+1}}{x_i} = -l_i^{-}, \qquad i = k, k+1, \ldots, n-1. \tag{16}$$

Taking (3) into account we derive recursive formulas for the rotation parameters c_i, s_i:

$$c_0 = 1, \qquad \frac{c_i}{s_i} = c_{i-1}\frac{x_{i+1}}{x_i}, \qquad i = 1, 2, \ldots, n-1. \tag{17}$$

Remark 2 *The ratios x_{i+1}/x_i must be computed by formulas (15), (16). It is not good to compute first the vector x by formulas (11), (12) and then the ratios of its neighboring components.*

7 Effects of the floating point numbers

Let us consider computations by the formulas

$$\begin{aligned} d_1^+ &= d_1 - \lambda, \qquad d_i^+ = d_i - \lambda - a_{i-1}u_{i-1}^+, \qquad i = 2, \ldots, k-1, \\ u_i^+ &= b_i/d_i^+, \qquad i = 1, \ldots, k-1. \end{aligned} \tag{18}$$

The formulas for l_i^+ are not used in the implementations.

If in the course of computations by formulas (18) some d_i takes zero value, then such d_i is replaced with a nonzero $d_{\min}$, which is chosen below.

Let us denote $\eta = \max_{i,j,m}\{|d_i|, |a_j|, |b_m|\}$. Owing to Gershgorin's theorem we assume $|\lambda| \le 3\eta$.

Following MATLAB's notation we designate the relative precision by ϵ, the maximum real number by *realmax*, and the minimum positive real number by *realmin*.

According to the theory of rounding errors

$$|fl(u_i^+)| \le \frac{\eta(1+\epsilon)}{d_{\min}}, \qquad |fl(d_i^+)| \le (1+\epsilon)^2 \left[4\eta + \eta^2(1+\epsilon)/d_{\min}\right].$$

If $\eta \le \epsilon \cdot realmax$, the restrictions $\eta/d_{\min} \le realmax/2$ and $\eta^2/d_{\min} \le realmax/2$ guarantee with abundance that no overflow will occur when computing u_i^+ and d_i^+. In addition, we require that $d_{\min} \ge realmin$. Taking all into account, we choose $d_{\min}$ equal to

$$d_{\min} = \epsilon_0 \max(1, \eta, \eta^2), \quad \text{where } \epsilon_0 = \max(realmin, 2/realmax).$$

Since with the replacement of $fl(d_i^+)$ by $d_{\min}$ we commit a relative error of size $d_{\min}$ in T, it is natural to impose a restriction $d_{\min}/\eta \le \epsilon$. If $\eta \le 1$, then $d_{\min} = \epsilon_0$ and we get a restriction on η of the form $\eta \ge \epsilon_0/\epsilon$. If $\eta > 1$, then $d_{\min} = \epsilon_0\eta^2$ and it follows the restriction $\eta \le \epsilon/\epsilon_0$, which immediately implies $\eta \le \epsilon \cdot realmax$.

Let us summarize the above discussion. We introduced the parameter

$$\epsilon_0 = \max\{realmin, 2/realmax\}. \tag{19}$$

If $\eta = \max_{i,j,m}\{|d_i|, |a_j|, |b_m|\}$ does not satisfy the bounds

$$\epsilon_0/\epsilon \le \eta \le \epsilon/\epsilon_0, \tag{20}$$

then one needs to scale the matrix T and λ by powers of 2 in order to fulfill (20). The parameter $d_{\min}$ is chosen as

$$d_{\min} = \epsilon_0 \max(1, \eta^2). \tag{21}$$

Computations by formulas (9) are carried out similarly as above.

Proceeding to the analysis of computations by the formulas

$$\begin{aligned} c_0 &= 1, \qquad \tfrac{c_i}{s_i} = c_{i-1}\left(-\tfrac{1}{u_i^+}\right), \qquad i = 1, 2, \ldots, k-1, \\ \tfrac{c_i}{s_i} &= c_{i-1}\left(-l_i^-\right), \qquad i = k, k+1, \ldots, n-1, \end{aligned} \tag{22}$$

note that the underflow in c_{i-1} is not acceptable. Therefore, we have to use the floating point numbers with so called extended exponents. It means that all arithmetical operations are executed with pairs (f, e), where $1/2 \le f < 1$ and e is a standard or long integer.

A question how to treat the cases where $fl(u_i^+)$ or $fl(l_i^-)$ equal zero is related to the latter discussion. If these happen, one has to substitute u_i^+ and l_i^- in (22) with b_i/d_i^+ and a_i/d_{i+1}^- respectively and again use the floating point numbers with extended exponents.

After computations with the above precautions the rotation parameters (before possible underflow caused by the conversion from the extended exponents to the standard ones) correspond to some tridiagonal matrix $\widehat{T}$ which is a perturbation of the original matrix T. If T is real symmetric, then the perturbation theory of Section 3 is applicable.

8 Implementation of the deflation in MATLAB

The MATLAB function DFLGEN presented below computes a chain of rotations which deflates a real symmetric tridiagonal matrix T with an approximate eigenvalue λ.

```
function [c,s] = dflgen(t,lambda)
%DFLGEN Generation of deflation parameters for a symmetric tridiagonal matrix.
%       [C,S] = DFLGEN(T,LAMBDA) computes parameters C and S of
%       a chain of N-1 planar rotations that deflates a real N-by-N
%       symmetric tridiagonal matrix T corresponding to an eigenvalue LAMBDA.
%       Only the lower triangular part of T is referenced.

n = size(t,1);

d  = diag(t); dl = diag(t,-1);

rmin = max(realmin, 2/realmax);

r = max(norm(d,inf), norm(dl,inf));
if r < rmin/eps | r > eps/rmin     % Scale T and LAMBDA by a power of 2
  [r,e] = log2(r);
  d  = pow2(d,-e); dl = pow2(dl,-e);
  lambda = pow2(lambda,-e);
end

d = d - lambda;
dl(dl==0) = realmin;               % Replace zero subdiagonals by REALMIN

dmin = max(rmin, (rmin*r)*r);      % Replacement for very small pivots
                                   % in LDU and UDL factorizations

% LDU factorization of T-lambda*I

d1 = d; u1 = zeros(n-1,1);
for i = 2:n
  u1(i-1) = dl(i-1)/d1(i-1);
  d1(i) = d1(i) - dl(i-1)*u1(i-1);
  if abs(d1(i)) < dmin
     d1(i) = dmin;
  end
```

```
end

% UDL factorization of T-lambda*I

d2 = d; l2 = zeros(n-1,1);
for i = n-1:-1:1
  l2(i) = dl(i)/d2(i+1);
  d2(i) = d2(i) - dl(i)*l2(i);
  if abs(d2(i)) < dmin
     d2(i) = dmin;
  end
end

[gamma,k] = min(abs(d1+d2-d));     % Optimal k and gamma_k

c = ones(n-1,1); s = zeros(n-1,1);

[cf,ce] = log2(1);                 % c_0 = 1 in the form
                                   % with extended exponents
for i = 1:n-1
  if i < k                         % sf*2^se = u1(i)
    if u1(i) ~= 0
      [sf,se] = log2(u1(i));
    else                           % Recompute u1(i) with extended exponents
      [sf,se] = log2(dl(i));
      [f,e] = log2(d1(i));
      sf = sf/f; se = se-e;
    end
  else                             % sf*2^se = 1/l2(i)
    if l2(i) ~= 0
      [f,e] = log2(l2(i));
      sf = 1/f; se = -e;
    else                           % Recompute 1/l2(i) with extended exponents
      [sf,se] = log2(d2(i+1));
      [f,e] = log2(dl(i));
      sf = sf/f; se = se-e;
    end
  end

% Convert [cf*2^ce, sf*2^se] into [c(i), -s(i)]

  e = max(ce,se); ce = ce-e; se = se-e;
  r = norm([pow2(cf,ce); pow2(sf,se)]);
  [cf,ce1] = log2(cf/r);           % Save c_i in the form
  ce = ce+ce1;                     % with extended exponents
  c(i) = pow2(cf,ce);
  s(i) = -pow2(sf/r,se);
end
```

The following MATLAB function computes the deflated tridiagonal matrix with the rotation parameters computed by DFLGEN.

```
function t1 = dflapp(t,c,s)
%DFLAPP Deflation of a real symmetric tridiagonal matrix T.
%       T1 = DFLAPP(T,C,S) computes a result of deflation of a real
%       symmetric tridiagonal matrix T by a chain of planar rotations
%       produced in DFLGEN.

t1 = t;
n = size(t1,1);

for i = 1:n-1
  G = [c(i) s(i); -s(i) c(i)];
  t1(i:i+1,:) = G'*t1(i:i+1,:);
  t1(:,i:i+1) = t1(:,i:i+1)*G;
end
```

9 Use of the deflation in the Lanczos process

The Lanczos process generates a decomposition

$$AU = UT_n + \beta_n u_{n+1} e_n^T, \tag{23}$$

where A is a large Hermitian matrix, T_n is a real symmetric n-by-n matrix, $[U, u_{n+1}]^*[U, u_{n+1}] = I_{n+1}$, $\beta_n \neq 0$ is a real constant. Assume that all subdiagonal elements of T_n are not zeros, i.e. T_n is unreduced.

Let us suppose that Ux is an approximate eigenvector of A corresponding to an approximate eigenvalue λ and $\|x\|_2 = 1$. This implies that the norm of the residual $AUx - \lambda Ux$ is tiny. Using the factorization (23) we conclude that the norm of the vector $U(T_n x - \lambda x) + \beta_n x_n u_{n+1}$ is small. Therefore, x is an approximate eigenvector of T_n corresponding to an approximate eigenvalue λ and $|\beta_n x_n|$ is tiny.

On the base of the above argument we will assume that

a) $T_n x = \lambda x$, where $x \in R^n$, $\|x\|_2 = 1$,

b) $|\beta_n x_n|$ is very small with respect to $\|A\|_2$.

Choose a chain of rotations G with real c_i, s_i so that $G^T x = \rho_n e_n$, $|\rho_n| = 1$. As a consequence of Kirilyuk's theorem,

$$G^T T_n G = \begin{pmatrix} T_{n-1} & 0 \\ 0 & \lambda \end{pmatrix} \tag{24}$$

with a symmetric tridiagonal matrix T_{n-1}. Plugging (24) into (23) we derive that

$$AUG = UG \begin{pmatrix} T_{n-1} & 0 \\ 0 & \lambda \end{pmatrix} + \beta_n u_{n+1} (G_{n-1} e_n)^T,$$

where $(G_{n-1} e_n)^T = [0 \ \dots \ 0 \; s_{n-1} \; c_{n-1}]$. With the help of the permutation matrix

$$P_n = \begin{pmatrix} 0 & I_{n-1} \\ 1 & 0 \end{pmatrix}.$$

we get the following equality:

$$A(UGP_n) = (UGP_n)\begin{pmatrix} \lambda & 0 \\ 0 & T_{n-1} \end{pmatrix} + \beta_n s_{n-1} u_{n+1} e_n^T + \beta_n c_{n-1} u_{n+1} e_1^T. \tag{25}$$

Since $|\beta_n x_n| = |\beta_n c_{n-1}|$, the last member in (25) can be omitted. As a result, we come at the final decomposition

$$A\widetilde{U} = \widetilde{U}\begin{pmatrix} \lambda & 0 \\ 0 & T_{n-1} \end{pmatrix} + \tilde{\beta}_n u_{n+1} e_n^T, \tag{26}$$

where $\tilde{\beta}_n = \beta_n s_{n-1}$, $\widetilde{U} = UGP_n$.

The above deflation scheme for the Lanczos process has one defect. Since the rotation parameters are computed from the matrix $T_n - \lambda I$ and not from the components of the vector x, there is no guarantee that the computed chain of rotations polarizes x rather than any other eigenvector with an eigenvalue close to λ. Therefore, we recommend to check the value of $|\beta_n c_{n-1}|$ instead of $|\beta_n x_n|$. However, this recommendation does not get rid of the defect.

The proposed deflation scheme is more efficient than the one suggested in [6].

10 An interior deflation

Assume that a real symmetric tridiagonal matrix

$$T = \begin{bmatrix} d_1 & b_1 & & & \\ b_1 & d_2 & b_2 & & \\ & \cdot & \cdot & \cdot & \\ & & b_{n-2} & d_{n-1} & b_{n-1} \\ & & & b_{n-1} & d_n \end{bmatrix}$$

has an approximate zero eigenvalue. Let us denote the QR factorization of T by Q_+R and QL factorization of T by Q_-L. Then

$$Q_+ = G_1^+ G_2^+ \dots G_{n-1}^+, \qquad Q_- = G_{n-1}^- G_{n-2}^- \dots G_1^-,$$

where

$$G_i^+ = \begin{pmatrix} I & & & \\ & c_i^+ & s_i^+ & \\ & -s_i^+ & c_i^+ & \\ & & & I \end{pmatrix}, \qquad G_i^- = \begin{pmatrix} I & & & \\ & c_i^- & s_i^- & \\ & -s_i^- & c_i^- & \\ & & & I \end{pmatrix}.$$

The blocks $\begin{pmatrix} c_i & s_i \\ -s_i & c_i \end{pmatrix}$ in both matrices G_i^+ and G_i^- stand at the intersection of rows i, $i+1$ and columns i, $i+1$.

The matrix R is computed as follows:

$$\begin{gathered} \widehat{R}_{11} = d_1, \quad \widehat{R}_{12} = b_1, \\ \begin{bmatrix} c_i^+ & -s_i^+ \\ s_i^+ & c_i^+ \end{bmatrix}\begin{bmatrix} \widehat{R}_{ii} & \widehat{R}_{i,i+1} & 0 \\ b_i & d_{i+1} & b_{i+1} \end{bmatrix} = \begin{bmatrix} R_{ii} & R_{i,i+1} & R_{i,i+2} \\ 0 & \widehat{R}_{i+1,i+1} & \widehat{R}_{i+1,i+2} \end{bmatrix}, \quad i = 1, \dots, n-2, \\ \begin{bmatrix} c_{n-1}^+ & -s_{n-1}^+ \\ s_{n-1}^+ & c_{n-1}^+ \end{bmatrix}\begin{bmatrix} \widehat{R}_{n-1,n-1} & \widehat{R}_{n-1,n} \\ b_{n-1} & d_n \end{bmatrix} = \begin{bmatrix} R_{n-1,n-1} & R_{n-1,n} \\ 0 & R_{nn} \end{bmatrix}. \end{gathered} \tag{27}$$

The matrix L is computed similarly:

$$
\begin{aligned}
&\widehat{L}_{n,n-1} = b_{n-1}, \quad \widehat{L}_{nn} = d_n, \\
&\begin{bmatrix} c_i^- & -s_i^- \\ s_i^- & c_i^- \end{bmatrix} \begin{bmatrix} b_{i-1} & d_i & b_i \\ 0 & \widehat{L}_{i+1,i} & \widehat{L}_{i+1,i+1} \end{bmatrix} = \begin{bmatrix} \widehat{L}_{i,i-1} & \widehat{L}_{i,i} & 0 \\ L_{i+1,i-1} & L_{i+1,i} & L_{i+1,i+1} \end{bmatrix}, \quad i = n-1, \dots, 2, \\
&\begin{bmatrix} c_1^- & -s_1^- \\ s_1^- & c_1^- \end{bmatrix} \begin{bmatrix} d_1 & b_1 \\ \widehat{L}_{21} & \widehat{L}_{22} \end{bmatrix} = \begin{bmatrix} L_{11} & 0 \\ L_{21} & L_{22} \end{bmatrix}.
\end{aligned}
\tag{28}
$$

Let us fix some k, $1 \le k \le n-1$, and define the matrices

$$
\begin{array}{rcll}
W_{n-1} & = & G_1^+ G_2^+ \dots G_{n-2}^+, & k = n-1, \\
W_k & = & G_1^+ G_2^+ \dots G_{k-1}^+ G_{n-1}^- G_{n-2}^- \dots G_{k+1}^-, & 1 < k < n-1, \\
W_1 & = & G_{n-1}^- G_{n-2}^- \dots G_2^-, & k = 1,
\end{array}
$$

then the matrix $W_k^T T$ has the following structure:

$$
W_k^T T = \begin{bmatrix}
R_{11} & R_{12} & R_{13} & & & & & & \\
 & \ddots & \ddots & \ddots & & & & & \\
 & & R_{k-1,k-1} & R_{k-1,k} & R_{k-1,k+1} & & & & \\
 & & & \widehat{R}_{kk} & \widehat{R}_{k,k+1} & & & & \\
 & & & \widehat{L}_{k+1,k} & \widehat{L}_{k+1,k+1} & & & & \\
 & & & L_{k+2,k} & L_{k+2,k+1} & L_{k+2,k+2} & & & \\
 & & & & \ddots & \ddots & \ddots & & \\
 & & & & & L_{n,n-2} & L_{n,n-1} & L_{nn} &
\end{bmatrix}.
$$

Choose a matrix $G_k = \begin{pmatrix} I & & & \\ & c_k & s_k & \\ & -s_k & c_k & \\ & & & I \end{pmatrix}$ such that

$$
\begin{bmatrix} c_k & -s_k \\ s_k & c_k \end{bmatrix} \begin{bmatrix} \widehat{R}_{kk} & \widehat{R}_{k,k+1} \\ \widehat{L}_{k+1,k} & \widehat{L}_{k+1,k+1} \end{bmatrix} = \begin{bmatrix} \widetilde{L}_{kk} & 0 \\ \widetilde{L}_{k+1,k} & \widetilde{L}_{k+1,k+1} \end{bmatrix}.
$$

Denote $Q_k = W_k G_k$ for $1 \le k \le n-1$ and $Q_n = Q_+$. Then one can easily prove that the matrix $T_{Q_k} = Q_k^T T Q_k$ has the following structure:

$$
T_{Q_k} = \begin{bmatrix}
\times & \times & & & & & & & & & \\
\times & \times & \times & & & & & & & & \\
 & \ddots & \ddots & \ddots & & & & & & & \\
 & & \times & \times & \times & & & & & & \\
 & & & \times & \times & \times & \times & \times & & & \\
 & & & & \times & \times & \times & \times & & & \\
 & & & & \times & \times & \times & \times & & & \\
 & & & & \times & \times & \times & \times & \times & & \\
 & & & & & & & \times & \times & \times & \\
 & & & & & & & & \ddots & \ddots & \ddots \\
 & & & & & & & & & \times & \times
\end{bmatrix}.
$$

It is a real tridiagonal symmetric matrix with a bulge at the intersection of rows $k-1$, k, $k+1$, $k+2$ and columns $k-1$, k, $k+1$, $k+2$.

Let us consider the equation $Tx = \gamma_k e_k$ with $x_k = 1$. The fundamental result from [2], [8] says that there exists k such that γ_k is tiny if 0 is a good approximate eigenvalue of T. The k-th equation of the system $Q_k^T T x = \gamma_k Q_k^T e_k$ is

$$\widetilde{L}_{kk} = \widetilde{L}_{kk} x_k = \gamma_k c_{k-1}^{+} c_k, \quad 1 \le k \le n-1,$$

or

$$R_{nn} = R_{nn} x_n = \gamma_n c_{n-1}^{+}, \quad k = n.$$

The existence of tiny γ_k implies the existence of a negligibly small $\widetilde{L}_{kk}$ or R_{nn}. So we choose k as an index of the smallest element in the array $\{|\widetilde{L}_{11}|, |\widetilde{L}_{22}|, \ldots, |\widetilde{L}_{n-1,n-1}|, |R_{nn}|\}$. The norm of the k-th row of T_{Q_k} is equal to the value of this element and therefore is negligibly small. As a result, the k-th row and k-th column of T_{Q_k} are approximately equal to 0.

We denote by Π_k the permutation matrix

$$\Pi_k = \begin{pmatrix} I_{k-1} & 0 & 0 \\ 0 & 0 & 1 \\ 0 & I_{n-k} & 0 \end{pmatrix}$$

that move the k-th column to the last position. The matrix $\Pi_k^T T_{Q_k} \Pi_k$ is a small perturbation of

$$\begin{pmatrix} T_d & 0 \\ 0 & 0 \end{pmatrix},$$

where T_d is symmetric tridiagonal with a bulge at entries $(k-1, k+1)$ and $(k+1, k-1)$. This bulge is chased by rotations down or up.

11 Implementation of the internal deflation in MATLAB

```
function [t1,k,c,s,cb,sb] = intdfl(t,lambda)
%INTDFL Interior deflation of a real symmetric tridiagonal matrix.
%       [T1,K,C,S,CB,SB] = INTDFL(T,LAMBDA) carries out the interior
%       deflation of a real symmetric tridiagonal N-by-N matrix T with
%       an approximate eigenvalue LAMBDA. The deflated matrix is stored
%       in T1, K indicates the row of deflation.
%       The cross-chain of rotations is given by parameters
%       C(1:K-1, N-1:-1:K) and S(1:K-1, N-1:-1:K).
%       The rotations given by CB(1:N-2) and SB(1:N-2) chase
%       a possible internal bulge down.

n = size(t,1);
crossblock = zeros(2,2,n-1);

% QR factorization of T-lambda*I

r  = t - lambda*eye(n,n);
cr = ones(n-1,1); sr = zeros(n-1,1);

for i = 1:n-1
```

```
  crossblock(1,:,i) = r(i,i:i+1);
  G = givens(r(i,i),-r(i+1,i));
  r(i:i+1,i:min(i+2,n)) = G'*r(i:i+1,i:min(i+2,n));
  cr(i) = G(1,1); sr(i) = G(1,2);
end

% QL factorization of T-lambda*I

l  = t - lambda*eye(n,n);
cl = ones(n-1,1); sl = zeros(n-1,1);

for i = n-1:-1:1
  crossblock(2,:,i) = l(i+1,i:i+1);
  G = givens(l(i+1,i+1),l(i,i+1));
  l(i:i+1,max(i-1,1):i+1) = G'*l(i:i+1,max(i-1,1):i+1);
  cl(i) = G(1,1); sl(i) = G(1,2);
end

% Triangularize the cross-blocks and determine an optimal k

p(n) = r(n,n);
c = ones(n-1,1); s = zeros(n-1,1);

for i = 1:n-1
  G = givens(crossblock(2,2,i),crossblock(1,2,i));
  crossblock(:,:,i) = G'*crossblock(:,:,i);
  p(i) = crossblock(1,1,i);
  c(i) = G(1,1); s(i) = G(1,2);
end

[pmin,k] = min(abs(p));

% Form the composite matrix and chain of deflating rotations

if k == n
  t1 = r;
  c = cr; s = sr;
else
  t1 = [r(1:k-1,1:n); zeros(2,n); l(k+2:n,1:n)];
  t1(k:k+1,k:k+1) = crossblock(:,:,k);
  c = [cr(1:k-1); c(k); cl(k+1:n-1)];
  s = [sr(1:k-1); s(k); sl(k+1:n-1)];
end

% Apply the deflating rotations from the right

for i = [1:k-1, n-1:-1:k]
   t1(:,i:i+1) = t1(:,i:i+1)*[c(i) s(i); -s(i) c(i)];
```

```
end

% Move the deflated eigenvalue to the bottom and chase the bulge down

t1 = t1([1:k-1 k+1:n k],[1:k-1 k+1:n k]) + lambda*eye(n,n);
cb = ones(n-2,1); sb = zeros(n-2,1);

if k > 1
  for i = k:n-2
    G = givens(t1(i,i-1),-t1(i+1,i-1));
    t1(i:i+1,i-1:i+2) = G'*t1(i:i+1,i-1:i+2);
    t1(i-1:i+2,i:i+1) = t1(i-1:i+2,i:i+1)*G;
    cb(i) = G(1,1); sb(i) = G(1,2);
  end
end
```

References

[1] I.S. Dhillon, *A new $O(n^2)$ algorithm for the symmetric tridiagonal eigenvalue/eigenvector problem*, PhD. thesis, University of California, Berkeley, 1997.

[2] K.V. Fernando, On computing an eigenvector of a tridiagonal matrix. Part I: Basic Results, *SIAM J. Matrix Anal. Appl.*, 18(4):1013–1034 (1997).

[3] S.K. Godunov, A.G. Antonov, O.P. Kiriljuk, and V.I. Kostin, *Guaranteed Accuracy in Numerical linear Algebra*, Kluwer Academic, 1993.

[4] S.K. Godunov, V.I. Kostin, and A.D. Mitchenko, *Computation of an eigenvector of a symmetric tridiagonal matrix*, Preprint 44, Institute of Math., Siberian Branch of the USSR Academy Sci., Novosibirsk, 58 pp., 1983. (In Russian)

[5] O.P. Kirilyuk, On transformations simplifying the structure of banded matrices, *Siberian Advances in Math.*, 1(3):1–25 (1991).

[6] R.B. Lehoucq, D.C. Sorensen, and C. Yang, *ARPACK User's Guide. Solution of Large-Scale Eigenvalue Problems with Implicitly Restarted Arnoldi Methods*, SIAM, 1998.

[7] A.N. Malyshev, *Introduction to Numerical Linear Algebra*, Nauka, Novosibirsk, Russia, 1991, 229 pp. (Russian)

[8] B.N. Parlett and I.S. Dhillon, Fernando's Solution to Wilkinson's Problem: An Application of double Factorization, *Linear Algebra and its Appl.* 267:247–279 (1997).

[9] J.H. Wilkinson, The calculation of eigenvectors of codiagonal matrices, *Computer J.*, 1:90–96 (1958).

[10] J.H. Wilkinson, *The Algebraic Eigenvalue Problem*, Clarendon Press, Oxford, 1965.