

# The Godunov–Inverse Iteration: A Fast and Accurate Solution to the Symmetric Tridiagonal Eigenvalue Problem

Anna M. Matsekh<sup>a,1</sup>

<sup>a</sup>*Institute of Computational Technologies, Siberian Branch of the Russian Academy of Sciences, Lavrentiev Ave. 6, Novosibirsk 630090, Russia*

---

## Abstract

We present a new hybrid algorithm based on Godunov’s method for computing eigenvectors of symmetric tridiagonal matrices and Inverse Iteration, which we call the *Godunov–Inverse Iteration*. We use eigenvectors computed according to Godunov’s method as starting vectors in the Inverse Iteration, replacing any non-numeric elements of Godunov’s eigenvectors with random uniform numbers. We use the right-hand bounds of the Ritz intervals found by the bisection method as Inverse Iteration shifts, while staying within guaranteed error bounds. In most test cases convergence is reached after only one step of the iteration, producing error estimates that are as good as or superior to those produced by standard Inverse Iteration implementations.

*Key words:* Symmetric eigenvalue problem, tridiagonal matrices, Inverse Iteration

---

## 1 Introduction

Construction of algorithms that enable to find all eigenvectors of the symmetric tridiagonal eigenvalue problem with guaranteed accuracy in  $O(n^2)$  arithmetic operations has become one of the most pressing problems of the modern numerical algebra. QR method, one of the most accurate methods for solving eigenvalue problems, requires  $6n^3$  arithmetic operations and  $O(n^2)$  square root operations to compute all eigenvectors of a tridiagonal matrix [Golub and

---

*Email address:* [matsekh@lanl.gov](mailto:matsekh@lanl.gov) (Anna M. Matsekh).

<sup>1</sup> Present address: Modeling, Algorithms and Informatics Group, Los Alamos National Laboratory. P.O. Box 1663, MS B256, Los Alamos, NM 87544, USA

Loan (1996)]. Existing implementations of the Inverse Iteration (e.g. LAPCK version of the Inverse Iteration xSTEIN [Anderson et al. (1995)] and EISPACK version of the Inverse Iteration TINVIT [Smith et al. (1976)], require  $O(n^2)$  floating point operations to find eigenvectors corresponding to well separated eigenvalues, but in order to achieve numerical orthogonality of eigenvector approximations corresponding to clustered eigenvalues, reorthogonalization procedures, such as Modified Gram-Schmidt process, are applied on each step of Inverse Iteration, increasing worst case complexity of the algorithm to  $O(n^3)$  operations. Typically Inverse Iteration for symmetric tridiagonal problem requires only three–five iteration steps [Wilkinson (1965), Smith et al. (1976), Anderson et al. (1995)] before convergence is achieved, but it tends to be very sensitive to the choice of the shift. If very accurate eigenvalue approximation is used as the Inverse Iteration shift convergence may not be achieved since shifted matrix in this case is nearly singular. To deal with this problem small perturbations are usually introduced in the shift parameter to avoid iteration breakdown, but as a rule the choice of such a perturbation is arbitrary, which may affect accuracy of the resulting eigenvectors.

Recently there have been a number of attempts to construct fast accurate algorithms that allow to compute eigenvectors corresponding to accurate eigenvalue approximations of tridiagonal symmetric matrices by dropping the redundant equation from the symmetric eigensystem. In 1995 K. V. Fernando proposed an algorithm for computing an accurate eigenvector approximation  $x$  from a nearly homogeneous system  $(A - \lambda I)x$  by finding an optimal index of the redundant equation that could be dropped [Fernando (1997)]. In 1997 Inderjit Dhillon proposed an  $O(n^2)$  algorithm for the symmetric tridiagonal eigenproblem [Dhillon (1997)] based on incomplete  $LDU$  and  $UDL$  (twisted) factorizations, which provided yet another solution to finding the redundant equation in a tridiagonal symmetric eigensystem. Much earlier, in 1983, S. K. Godunov and his collaborators [Godunov et al. (1988), Godunov et al. (1993)] proposed a two-sided Sturm sequence based method that enables to determine all eigenvectors of tridiagonal symmetric matrices with guaranteed accuracy in  $O(n^2)$  floating point operations by eliminating the redundant equation.

The use of the two-sided Sturm sequences in the Godunov’s method allows to avoid accuracy loss associated with rounding errors in the conventional Sturm sequence based methods. The algorithm gives provably accurate solutions to the symmetric tridiagonal eigenvalue problem on a specially designed floating point arithmetics with extended precision and directed rounding [Godunov et al. (1988)]. Unfortunately in IEEE arithmetics division by zero and overflow errors in the Godunov’s eigenvector computations can not be entirely avoided, while for computationally coincident and closely clustered eigenvalues it produces nearly collinear eigenvectors, taking no measures for reorthogonalization. In empirical studies Godunov’s method, direct method by it’s nature, consistently delivers residuals that are approximately two orders of magni-

tude larger than those of the eigenvectors computed according to some of the Inverse Iteration implementations. Inverse Iteration on the other hand often suffers from the nondeterministic character of starting vectors, the need to introduce disturbances into the shift, and the high worst case complexity.

In the attempt to overcome shortcomings of both Godunov's method and the Inverse Iteration we constructed a new hybrid procedure for computing eigenvectors of symmetric tridiagonal unreduced matrices which we call Godunov–Inverse Iteration. Godunov–Inverse Iteration can be viewed as Godunov's method with iterative improvement. It uses eigenvectors computed from the two-sided Sturm sequences as starting vectors. One step of the Inverse Iteration is usually sufficient to get desirable accuracy and orthogonality of the eigenvectors. This is in contrast to LAPACK version of the Inverse Iteration xSTEIN [Anderson et al. (1995)] which uses randomly generated starting vectors and requires three–five steps for convergence and EISPACK version of the Inverse Iteration TINVIT [Smith et al. (1976)] which uses direct solution to the tridiagonal problem as a starting vector in the Inverse Iteration and requires up to five steps for convergence. By choosing right-hand bounds of the the smallest machine presentable eigenvalue intervals, found by the bisection algorithm, as shifts in the Godunov–Inverse Iteration, instead of the eigenvalue approximation (the center of this interval) we insure that iteration matrix won't be numerically singular, and the perturbation does not exceed error bounds for the corresponding eigenvalue.

This paper is organized as follows. In section 2 we give a formal description of the symmetric eigenvalue problem. In section 3 we give an overview of the original version of the Godunov's method. In section 4 we discuss shortcomings of the Godunov's method and present Godunov–Inverse Iteration. We implemented and tested Godunov-Inverse Iteration, Godunov's method, Inverse Iteration with random starting vectors (LAPACK approach) and Inverse Iteration with starting vectors found as direct solutions to the eigenproblem (EISPACK approach), as well as Sturm sequence based bisection method, and Householder and Lanczos tridiagonalization procedures for dense and sparse matrices respectively. In section 5 we compare the quality of the eigenvectors computed with these methods on a tridiagonal, dense and sparse test matrices. Throughout the paper we assume that symmetric eigenvalue problems with matrices having arbitrary structure can be reduced to tridiagonal form with orthogonal transformations, which preserve spectral properties of original matrices to machine precision [Golub and Loan (1996)].

## 2 Formulation of the Problem

Consider the fundamental algebraic eigenvalue problem, in which

$$Ax = \lambda x \tag{1}$$

for real symmetric matrices  $A \in \mathbb{R}^{n \times n}$ . There always exists a real orthogonal transformation  $W \in \mathbb{R}^{n \times n}$  such that matrix  $A$  is diagonalizable [Wilkinson (1965)], that is,

$$W^T A W = \text{diag}(\lambda_i), \tag{2}$$

where eigenvalues  $\lambda_i$ ,  $i = 1, \dots, n$  are all real. We solve problem (1) with a real symmetric matrix  $A$  with arbitrary structure according to the Rayleigh–Ritz procedure [Godunov et al. (1988), Godunov et al. (1993)] that is, we

- (i) compute orthonormal transformation  $Q$  such that matrix  $T = Q^T A Q$  is tridiagonal
- (ii) solve eigenproblem  $Tu = \mu u$
- (iii) take  $(\mu, Qu)$  as an approximation to the eigenpair  $(\lambda, x)$

## 3 Godunov’s Method

Godunov’s method [Godunov et al. (1988), Godunov et al. (1993)] was designed to compute eigenvectors of an unreduced symmetric tridiagonal matrix

$$T = \begin{pmatrix} d_0 & b_0 & & \cdots & 0 \\ b_0 & d_1 & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & b_{n-2} \\ 0 & \cdots & & b_{n-2} & d_{n-1} \end{pmatrix} \tag{3}$$

by a sequence of plane rotations on a specially designed architecture that supports extended precision and directed rounding. Let  $(\alpha_i, \beta_i)$  be an eigeninterval that is guaranteed to contain an eigenvalue  $\mu_i(T)$  of the matrix  $T$ . Such an interval can be found by the bisection method with the accuracy [Godunov et al. (1988)]

$$|\beta_i - \alpha_i| \leq 3 \frac{(5\gamma + 1) \epsilon_{mach}}{2\gamma - (5\gamma + 1) \epsilon_{mach}} \mathfrak{M}(T), \tag{4}$$

where  $\gamma$  is the base used for the floating point exponent, and

$$\mathfrak{M}(T) = \max \begin{cases} |d_0| + |b_0| \\ \max_{1 \leq i < n-1} |d_i| + |b_i| + |b_{i+1}| \\ |d_{n-1}| + |b_{n-1}| \end{cases} . \quad (5)$$

Note that using Cauchy-Bunyakovsky inequality it can be established that  $\mathfrak{M}(T) \leq \sqrt{n} \|T\|_2$  [Godunov et al. (1988)]. Godunov eigenvector approximation  $u^i$  corresponding to the eigenvalue approximation  $\mu_i(T) \in (\alpha_i, \beta_i)$  is found recursively from the two-sided Sturm sequence  $P_k(\mu_i)$ ,  $i, k = 1, \dots, n$ , by letting

$$u_0^i = 1 \text{ and } u_k^i = -u_{k-1}^i \text{sign}(b_{k-1})/P_{k-1}(\mu_i) \quad (6)$$

in just  $O(n)$  operations per a normalized eigenvector. Two-sided Sturm sequence

$$P_0(\mu_i), \dots, P_{n-1}(\mu_i) \stackrel{\text{def}}{=} P_0^+(\alpha_i), \dots, P_l^+(\alpha_i), P_{l+1}^-(\beta_i), \dots, P_{n-1}^-(\beta_i) \quad (7)$$

is constructed from the left-sided and right-sided Sturm sequences  $P_k^+(\alpha_i)$ ,  $k = 0, \dots, n-1$  and  $P_k^-(\beta_i)$ ,  $k = n-1, \dots, 0$ . Left-sided Sturm sequence  $P_k^+(\alpha_i)$  is computed from the minors of the matrix  $T - \alpha_i I$  according to the formulas [Godunov et al. (1988)]

$$P_0^+(\alpha_i) = |b_0|/(d_0 - \alpha_i) \quad (8a)$$

$$P_k^+(\alpha_i) = |b_k|/(d_k - \alpha_i - |b_{k-1}|)P_{k-1}^+(\alpha_i) \quad (8b)$$

$$P_{n-1}^+(\alpha_i) = 1/(d_{n-1} - \alpha_i - |b_{n-2}|)P_{n-2}^+(\alpha_i) \quad (8c)$$

while right-sided Sturm sequence  $P_k^-(\beta_i)$  is computed from the minors of the matrix  $T - \beta_i I$  as follows [Godunov et al. (1988)]:

$$P_{n-1}^-(\beta_i) = d_{n-1} - \beta_i \quad (9a)$$

$$P_k^-(\beta_i) = (d_k - \beta_i - |b_k|/P_{k+1}^-(\beta_i))/|b_{k-1}| \quad (9b)$$

$$P_0^-(\beta_i) = d_0 - \beta_i - |b_0|/P_1^-(\beta_i) \quad (9c)$$

Although analytically equivalent, in finite precision eigenvectors constructed from the left-sided and the right-sided Sturm sequences for the same parameter  $\lambda$  in general are different. An eigenvector with guaranteed accuracy is obtained when left and right hand sequences (7) are joint at an index  $l$  chosen according to the rule based on the Sturm theorem: for any real  $\lambda_0$  the number of roots  $\lambda$  of the  $k$ -th principal minor of the matrix  $T - \lambda I$ , such that  $\lambda < \lambda_0$  coincides with the number of nonpositive values in the Sturm sequence  $P(\lambda_i)_k$ ,  $k = 1, \dots, n$ . Let  $l^+$  be the number of nonpositive elements in the sequence  $P_k^+(\alpha_i)$ ,  $k = 0, \dots, n-1$ , and  $n-1-l^-$  be the number of nonnegative elements in the sequence  $P_k^-(\beta_i)$ ,  $k = n-1, \dots, 0$ . Then left and right sequences (7) are joint

at the index  $l = l^+ = l^-$  for which the following condition is satisfied [Godunov et al. (1988)]:

$$(P_l^+(\alpha_i) - P_l^-(\beta_i))(1/P_{l+1}^-(\beta_i) - 1/P_{l+1}^+(\alpha_i)) \leq 0 \quad (10)$$

Resulting two-sided Sturm sequence (7) is used to recursively compute Godunov eigenvectors (6).

#### 4 Godunov–Inverse Iteration

In our attempt to improve Godunov’s method we were motivated by the fact that it is a direct method, and due to the rounding errors in finite precision theoretical error bound for the eigenvectors computed according to the Godunov’s method [Godunov et al. (1988)]:

$$\|(T - \mu_k I)u_k\|_2 \leq 13\sqrt{3} \epsilon_{mach} \|T\|_2 \|u_k\|_2 \quad (11)$$

is not achieved. At the same time two-sided Sturm sequence computations are susceptible to division by zero and overflow errors, while for closely clustered and computationally coincident eigenvalues it produces nearly collinear eigenvectors, taking no measures for reorthogonalization. In empirical studies it consistently delivered residuals that were approximately two orders of magnitude larger than those of the eigenvectors computed by Inverse Iteration with random starting vectors. In addition, due to the rounding errors, machine representation of the matrix  $T$ , that is generally obtained by either Householder or Lanczos tridiagonalization, has the form  $T_{mach} = T + G$  [Wilkinson (1965)], where

$$\|G\|_2 \leq k\sqrt{n}2^{-t} \quad (12)$$

and  $t$  is the number of mantissa bits in the machine representation of floating-point numbers. Therefore in finite precision error bound (11) rather takes the following form:

$$\|(T_{mach} - \mu_k I)u_k\|_2 \leq 13\sqrt{3} \epsilon_{mach} \|T\|_2 \|u_k\|_2 + k\sqrt{n}2^{-t} \|u_k\|_2. \quad (13)$$

We constructed Godunov–Inverse Iteration to avoid common computational problems arising in both Godunov’s and Inverse Iteration methods. It can be viewed as an algorithm that delivers reorthogonalized iteratively improved Godunov eigenvectors. Instead of initiating Inverse Iteration with a random vector, or solving a linear system to find a starting vector, as it is customary in the implementations of the Inverse Iteration, we use Godunov’s eigenvector, computed in just  $O(n)$  arithmetic operations as an extremely accurate starting vector in the Inverse Iteration. Before Inverse Iteration is applied, any non-numeric elements of the Godunov’s eigenvectors are substituted with random

numbers. This semi-deterministic approach to finding initial vectors to the Inverse Iteration reduces the number of steps necessary for convergence to desired accuracy. In most cases convergence is achieved after one step of Inverse Iteration.

Inverse Iteration may break down when very accurate eigenvalue approximations  $\lambda$  are used as shifts, because shifted iteration matrix

$$A - \lambda I, \tag{14}$$

in this case is nearly singular (here  $I$  is an identity matrix). To avoid this, small perturbations are usually introduced into the shift  $\lambda$  to assure convergence to the corresponding Ritz vectors. But even small arbitrary deviations of the Ritz values from exact eigenvalues may produce significant deviations of Ritz vectors from the actual eigenvectors. We solve this problem by using the right-hand bounds  $\beta_k$  of the eigenvalue intervals

$$\mu_i \in (\alpha_i, \beta_i), \quad i = 1, \dots, n \tag{15}$$

found by the bisection algorithm (either Sturm based [Godunov et al. (1988)] or inertia-based [Fernando (1998)] versions of the bisection algorithm) as accurate shifts that are guaranteed to be within the error bounds (4). We apply Modified Gram–Schmidt reorthogonalization for the eigenvector approximations corresponding to computationally multiple eigenvalues or to the clustered eigenvalues with small relative gaps. We use Wilkinson’s stopping criteria [Wilkinson (1965)]

$$\|x_k\|_\infty \geq 2^t/100n \tag{16}$$

to verify that convergence is achieved. Below we present a formal description of the Godunov–Inverse Iteration.

**Algorithm 1 (Godunov-Inverse Iteration)**

Compute eigenvectors  $u^i$ ,  $i = 1, \dots, n$  of the tridiagonal matrix  $T = T^T \in \mathbb{R}^{n \times n}$  with main diagonal  $d$  and codiagonal  $b$  on an processor with machine precision  $\epsilon_{mach}$  and  $t$  mantissa bits.

**godunov\_inverse\_iteration(d, b, n)**

**bisection(d, b, n)**

**for** ( $i = 1, i \leq n, i++$ )

    find eigenintervals  $(\alpha_i, \beta_i)$  that contain eigenvalues  $\mu_i \in (\alpha_i, \beta_i)$  s.t.

$|\beta_i - \alpha_i| \leq \epsilon_{mach}(|\beta_i| + |\alpha_i|)$ ,  $i = 1, \dots, n$

**end**

**godunov\_eigenvector\_method(d, b, n,  $\alpha$ ,  $\beta$ )**

**for** ( $i = 1, i \leq n, i++$ )

    compute two-sided Sturm sequence  $P_n(\mu_i)$ :

$$\begin{aligned}
P_1^+(\alpha_i) &= |b_1|/(d_1 - \alpha_i) \\
P_k^+(\alpha_i) &= |b_k|/(d_k - \alpha_i - |b_{k-1}|)P_{k-1}^+(\alpha_i), \quad k = 2, 3, \dots, n-1 \\
P_n^+(\alpha_i) &= 1/(d_n - \alpha_i - |b_{n-1}|)P_{n-1}^+(\alpha_i)
\end{aligned}$$

$$\begin{aligned}
P_n^-(\beta_i) &= d_n - \beta_i \\
P_k^-(\beta_i) &= (d_k - \beta_i - |b_k|/P_{k+1}^-(\beta_i))/|b_{k-1}|, \quad k = n-1, n-2, \dots, 2 \\
P_1^-(\beta_i) &= d_1 - \beta_i - |b_1|/P_2^-(\beta_i)
\end{aligned}$$

find  $l = l^+ = l^-$  s.t.  $(P_l^+(\alpha_i) - P_l^-(\beta_i))(1/P_{l+1}^-(\beta_i) - 1/P_{l+1}^+(\alpha_i)) \leq 0$

set  $P_1(\mu_i), \dots, P_n(\mu_i) \stackrel{\text{def}}{=} P_1^+(\alpha_i), \dots, P_l^+(\alpha_i), P_{l+1}^-(\beta_i), \dots, P_n^-(\beta_i)$

compute Godunov's eigenvector  $u^i$ :  $u_1^i = 1$ ,

**for** ( $k = 2, k \leq n, k++$ )

$$u_k^i = -u_{k-1}^i \text{sign}(b_{i-1})/P_{i-1}(\mu_i)$$

**end**

**end**

**inverse\_iteration(d, b, u,  $\beta$ )**

preprocessing

**for** ( $i = 1, i \leq n, i++$ )

**for** ( $k = 1, k \leq n, k++$ )

**if**  $u_k^i$  is not a machine number

**then** set  $u_k^i$  to a random uniform number from  $(0, 1)$

**end**

use right ends of the eigenintervals as shifts:  $\gamma_i = \beta_i$

perturb clustered and computationally coincident eigenvalues:

**if** ( $i > 0 \cap |\gamma_i - \gamma_{i-1}| \leq 10 \epsilon_{mach} |\gamma_i|$ )

**then**  $\gamma_i = \gamma_{i-1} + 10\epsilon_{mach} |\gamma_i|$

**end**

inverse iteration

$k = 0, \delta = 2^t/(100 * n)$

**do**

$k = k + 1$

Solve  $(T - \gamma_k I)z = u^k$

**for** ( $j = 1, j < k, j++$ )

reorthogonalize  $z$  if it is almost collinear to an eigenvector in the basis:

**if**  $|\gamma_j - \gamma_k| \leq \mathfrak{M}(T)/1000$

**then**  $z = z - (z, u^j)u^j$

**end**

$u^k = z/\|z\|_2$

**while** ( $\|z\|_\infty > \delta$ )

**end**

## 5 Experimental Results

We implemented and tested Godunov’s method, Godunov–Inverse Iteration (Algorithm 1), Inverse Iteration algorithm with random starting vectors which we call Random Inverse Iteration algorithm (our implementation of the LAPACK procedure xSTEIN [Anderson et al. (1995)]), and Inverse Iteration algorithm with initial vectors found as a direct solution to the eigenproblem, which we call Direct Inverse Iteration algorithm (our implementation of the EISPACK procedure TINVIT [Smith et al. (1976)]), in ANSI C (GNU C compiler) in IEEE double precision and tested these programs on an Intel® Xeon™ CPU 1500MHz processor.

To make fair comparison we compute eigenvalue approximations only once and use these eigenvalues to compute eigenvectors using four different routines, while in all three Inverse Iteration implementations we use the same direct solver for systems of linear algebraic equations with tridiagonal symmetric matrices. We use Householder tridiagonalization with dense matrices and restarted Lanczos procedure with selective reorthogonalization with sparse matrices. Following Godunov [Godunov et al. (1988)] we implemented Wilkinson’s Sturm sequence based bisection algorithm [Wilkinson (1965)] to find intervals  $(\alpha_i, \beta_i)$  containing eigenvalues  $\mu_i$  of the tridiagonal matrix  $T = Q^T A Q$  with guaranteed accuracy

$$|\beta_i - \alpha_i| \leq \epsilon_{mach} \mathfrak{F}(T), \quad i = 1, \dots, n, \quad (17)$$

where  $\epsilon_{mach}$  is the unit roundoff error and  $\mathfrak{F}(T)$  is a constant that typically is a function of some norm of the matrix  $T$ . We generally terminate bisection iteration when the following condition is satisfied [Golub and Loan (1996)]:

$$|\beta_i - \alpha_i| \leq \epsilon_{mach} (|\beta_i| + |\alpha_i|), \quad i = 1, \dots, n \quad (18)$$

or the number of iterations equals  $t$  – the number of bits of precision in the machine representation of doubles. Indeed in the bisection method we are locating intervals  $(\beta_k, \alpha_k)$  of width  $(\beta_0 - \alpha_0)/2^t$  in  $t$  steps [Wilkinson (1965)]. After  $t = 52$  iterative steps in IEEE double precision the width of the interval is comparable to  $\epsilon_{mach} = 2^{-52}$  which is roughly  $2.22e - 16$ . As a result bisection algorithm requires  $O(t n^2)$  operations to compute all eigenvalues of a tridiagonal symmetric matrix. Intervals  $(\alpha_i, \beta_i)$  were used in the original Godunov’s method and in the new Godunov–Inverse Iteration, while in the Random Inverse Iteration and Direct Inverse Iteration versions of the algorithm  $\mu_i = (\alpha_i + \beta_i)/2$  is used as the  $i$ -th eigenvalue approximation. By using criterion (18) we get slightly smaller eigenintervals, and as a result slightly more accurate eigenvalue approximations than by using the following stop-

ping criterion (Godunov's criterion (4) multiplied by 1/3):

$$|\beta_i - \alpha_i| \leq \frac{(5\gamma + 1) \epsilon_{mach}}{2\gamma - (5\gamma + 1) \epsilon_{mach}} \mathfrak{M}(T). \quad (19)$$

In the Test Example 1 (Tables 1 and 2) we show that when we use eigenvalue approximations computed using criterion (18), corresponding eigenvectors computed with algorithms used in the LAPACK and EISPACK versions of Inverse Iteration may not converge, while eigenvalue approximations computed using criterion (19) to slightly lower accuracy ensured convergence of these algorithms to the desired eigenvectors. Godunov's and Godunov-Inverse Iteration methods appeared to be robust to the choice of the bisection stopping criterion.

We report the accuracy of the computed eigenpairs  $(\tilde{\lambda}_i, \tilde{x}_i)$ ,  $i = 1, 2, \dots, n$  of the matrix  $A = QTQ^T$  by testing whether the maximum residual vector norm normalized by the spectral radius of the matrix  $A$

$$\max_i \|(A - \tilde{\lambda}_i I) \tilde{x}_i\|_\infty / \max_i |\tilde{\lambda}_i| \quad (20)$$

is in the order of the unit roundoff error  $\epsilon_{mach}$ , and that the maximum deviation of the computed eigenvectors  $x_i$ ,  $i = 1, 2, \dots, n$ , from the unit vectors  $e_i$ ,  $i = 1, 2, \dots, n$ ,

$$\max_i \|(\tilde{X}^T \tilde{X} - I) e_i\|_\infty, \quad (21)$$

where  $\tilde{X} = |\tilde{x}_i|$ ,  $i = 1, 2, \dots, n$ , and  $I = |e_i|$ ,  $i = 1, 2, \dots, n$  is also in the order of the unit roundoff error  $\epsilon_{mach}$ . Indeed, in finite precision the accuracy of the computed eigenpairs  $(\tilde{\lambda}_i, \tilde{x}_i)$  of a matrix  $A$  is considered acceptable if the residual error norm is in the order of  $\epsilon_{mach} \|A\|$  [Godunov et al. (1988)]

$$\|(A - \tilde{\lambda}_i I) \tilde{x}_i\| = O(\epsilon_{mach} \|A\|), \quad (22)$$

while the orthonormality of the computed eigenbasis is expected to be in the order of  $\epsilon_{mach}$  [Godunov et al. (1988)]

$$\|\tilde{X}^T \tilde{X} - I\| = O(\epsilon_{mach}) \quad (23)$$

in a matrix norm typically induced by either infinity or euclidian vector norm. In all of the tests presented below Godunov-Inverse Iteration converged to desired accuracy in just one step, while the results were at the least as accurate as the ones obtained with the Random Inverse Iteration and Direct Inverse Iteration algorithms.

**Test Example 1** *Tridiagonal symmetric eigenproblem*  $Rx = \lambda x$ ,  $\lambda(R) = -\cos k\pi/(n+1)$ ,  $k = 1, \dots, n$ .

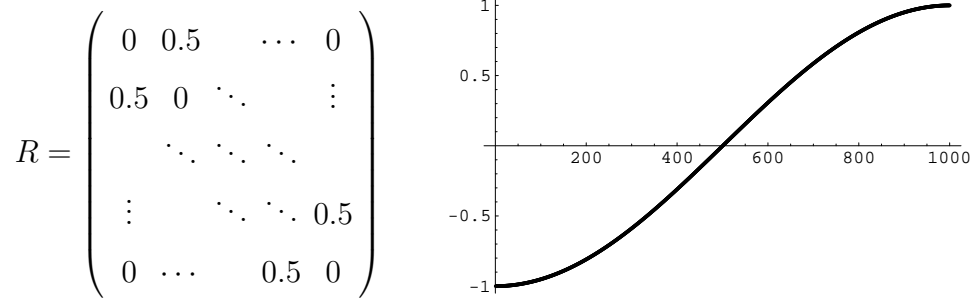


Fig. 1. Matrix  $R$  and its eigenvalues  $\tilde{\lambda}(R)$  computed by the bisection method for  $n = 1000$ .

	$\max_i \ (R - \tilde{\lambda}_i I)\tilde{x}_i\ _\infty / \max_i  \tilde{\lambda}_i $	$\max_i \ (\tilde{X}^T \tilde{X} - I)e_i\ _\infty$	#iters
Godunov's Method	$1.1535e - 12$	$2.9458e - 11$	-
Direct Inverse Iteration	$4.3393e - 12$	$1.0000e + 00$	1
Random Inverse Iteration	$2.3845e - 16$	$1.0000e + 00$	3
Godunov - Inverse Iteration	$2.3461e - 16$	$1.1138e - 14$	1

Table 1

Error estimates of the eigenvectors  $X = |x_k|_{k=1,\dots,n}$  of  $R$ , corresponding to the eigenvalues  $\tilde{\lambda}$  computed with maximum absolute deviation  $\Delta(\lambda) = 3.3307e - 16$  from the exact eigenvalues  $\lambda$  for  $n = 1000$ .

	$\max_i \ (R - \tilde{\lambda}_i I)\tilde{x}_i\ _\infty / \max_i  \tilde{\lambda}_i $	$\max_i \ (\tilde{X}^T \tilde{X} - I)e_i\ _\infty$	#iters
Godunov's Method	$2.0175e - 12$	$4.8817e - 11$	-
Direct Inverse Iteration	$5.9214e - 12$	$8.2479e - 11$	1
Random Inverse Iteration	$2.7557e - 16$	$2.2042e - 14$	3
Godunov - Inverse Iteration	$2.6822e - 16$	$1.0969e - 14$	1

Table 2

Error estimates of the eigenvectors  $X = |x_k|_{k=1,\dots,n}$  of  $R$ , corresponding to the eigenvalues  $\tilde{\lambda}$  computed with maximum absolute deviation  $\Delta(\lambda) = 4.9960e - 16$  from the exact eigenvalues  $\lambda$  for  $n = 1000$ .

The problem of finding eigenvalues and eigenvectors of tridiagonal symmetric matrices with zero main diagonal, in the so called Golub–Kahan form [Fernando (1998)], which arise in singular value computations for bidiagonal matrices, and generally for nonsymmetric matrices, presents a number of computational challenges. In the Test Example 1 we compare eigenvectors computed with the Godunov-Inverse Iteration against eigenvectors computed according to Godunov's method, and Direct and Random Inverse Iteration algorithms for the same approximations of the eigenvalues of the  $1000 \times 1000$  tridiagonal matrix  $R$  which has zero diagonal elements and elements equal 0.5 on the codiagonals. Eigenvalues of this matrix coincide with zeros of Chebyshev polynomials of second kind, and so we were able to compare analytical solution

against eigenvalues computed with our bisection routine. Test results for this example are summarized in the Tables 1 and 2.

when we use eigenvalue approximations computed using criterion (18) corresponding eigenvectors computed with algorithms used in the LAPACK and EISPACK versions of Inverse Iteration may not converge, while eigenvalue approximations computed with using Godunov’s criterion (19) to slightly lower accuracy ensure convergence of these algorithms to the desired eigenvectors.

For the eigenvalue approximations computed with maximum absolute deviation of  $3.3307e - 16$  (Table 1) from the analytical solution using criterion (18) some of the Direct Inverse Iteration and Random Inverse Iteration eigenvectors  $\tilde{X}$  did not converge and were set to zero, which is indicated by the fact that the maximum deviation from orthogonality equals 1 in these tests, yet Godunov’s eigenvectors satisfied orthogonality measure to 11 digits of machine precision. Godunov’s method and Direct Inverse Iteration produced eigenpairs that delivered residual errors that were accurate only to 12 digits of machine precision. In just one step of iterative improvement Godunov-Inverse Iteration produced eigenvectors that satisfied original problem to 16 digits of machine precision, just as Random Inverse Iteration solution did after three iteration steps. In addition Godunov-Inverse Iteration eigenvectors were orthonormal to at least 14 digits of machine precision.

When eigenvalues were computed with slightly lower precision (computed using criterion (4) with maximum absolute deviation of  $4.9960e - 16$  from the analytical solution; results presented in the Table 2) all three versions of the Inverse Iteration converged to high accuracy, and again Godunov-Inverse Iteration converged in only one step to virtually the same high accuracy as Random Inverse Iteration in three steps. Clearly traditional Inverse Iteration implementations appear to be very sensitive to the accuracy with which eigenvalue approximations are computed, while both Godunov’s method and Godunov-Inverse Iteration exhibit robust behavior.

**Test Example 2** *Dense symmetric eigenproblem  $Ux = \lambda x$ .*

$$U_{ij} = \begin{cases} 1/(i + j - 1) & i = j \\ -1/(i + j - 1) & i \neq j \end{cases}$$

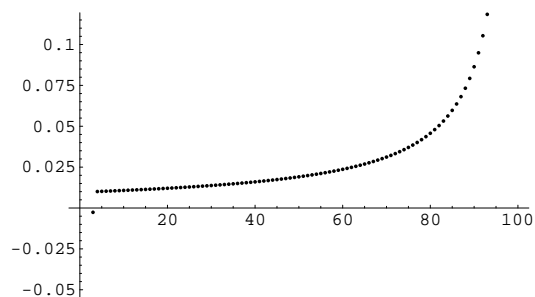


Fig. 2. Matrix  $U$  and its eigenvalues  $\tilde{\lambda}(U)$  computed by the bisection method for  $n = 100$ .



	$n = 225, c_2 = 1$	$n = 400, c_2 = 0.5$
$\Delta(\lambda)$	$6.6613e - 16$	$6.6613e - 16$
$\max_i \ (P - \tilde{\lambda}_i I)\tilde{x}_i\ _\infty / \max_i  \tilde{\lambda}_i $	$4.6359e - 16$	$5.2595e - 16$
$\max_i \ (\tilde{X}^T \tilde{X} - I)e_i\ _\infty$	$3.2918e - 15$	$4.8720e - 15$

Table 4

Error estimates of the eigenvectors  $X = |x_k|_{k=1,\dots,n}$  of  $P$  with  $c_0 = -0.33, c_1 = -0.17$ , corresponding to the eigenvalues  $\tilde{\lambda}$  computed with maximum absolute deviation  $\Delta(\lambda)$  from the exact eigenvalues  $\lambda$  by Godunov-Inverse Iteration method.

formed to the equivalent tridiagonal problem by Lanczos method with selective reorthogonalization. Matrix  $P \in \mathbb{R}^{n \times n}$ ,  $n = m^2$  is a version of the matrices arising in the finite difference approximations of the Laplacian on a rectangle. It has tridiagonal  $m \times m$  blocks on the main diagonal and codiagonals located  $m$  columns and  $m$  rows apart from the main diagonal. This is a very common test example with a known analytical solution and makes a good illustration of the correctness of our routines. Test results for this example are summarized in the Table 4: For  $n = 225$ ,  $c_0 = -0.33$ ,  $c_1 = -0.17$ ,  $c_2 = 1$  residual error was in the order of  $10^{-16}$  and the orthonormality error was in the order of  $10^{-15}$ . For  $n = 400$ ,  $c_0 = -0.33$ ,  $c_1 = -0.17$ ,  $c_2 = 0.5$  again residual error was in the order of  $10^{-16}$  and the orthogonality error was in the order of  $10^{-15}$ .

## 6 Conclusions

Godunov's method for real symmetric matrices produces accurate eigenvector approximations, but usually these vectors have fewer digits of precision than eigenvectors computed according to some of the Inverse Iteration implementations. Designed for unreduced matrices for computations on a specially designed architecture, in IEEE arithmetics in double precision Godunov's method produces almost collinear eigenvectors corresponding to closely clustered eigenvalues, and may even produce non-numeric output. At the same time the choices of the initial vector in the Inverse Iteration algorithms do not guarantee that starting vector has a nontrivial component in the direction of the solution, and the algorithms do not always converge. Inverse Iteration is very sensitive to the accuracy of the shift – we show that for eigenvalues computed by the bisection method with guaranteed accuracy in the order of machine precision Inverse Iteration algorithms used in the LAPACK in EISPACK packages may break down.

Godunov–Inverse Iteration was designed to solve these problems. Changing any non-numeric components of the Godunov eigenvectors to random uniformly distributed numbers, we apply Inverse Iteration to these vectors, which usually achieve desired error bounds in one step, in contrast with other im-

plementations of the Inverse Iteration algorithm which require a few more steps to achieve the same accuracy. This is most advantageous in the case of closely clustered eigenvalues when large fraction of the eigenvectors has to be reorthogonalized. Godunov–Inverse Iteration is very robust with respect to the choice of the Inverse Iteration shift – we use right-hand bounds of the eigenvalue intervals computed by the bisection method as extremely accurate shifts in the Godunov-Inverse Iteration. We resort to reorthogonalization within the iteration only in cases of computationally coincident or closely clustered eigenvalues. As a result Godunov-Inverse Iteration Algorithm produces accurate and robust solutions to the symmetric eigenvalue problem with higher accuracy than Godunov’s method and in fewer steps than existing implementations of the Inverse Iteration algorithm.

## 7 Acknowledgments

I’d like to thank Professor Sergei Konstantinovich Godunov for an insightful discussion of this project. I would also like to thank my adviser Professor Ella Petrovna Shurina, and my team leader at LANL Michael Pernice for valuable comments on drafts of the paper.

## References

- Anderson, E., Bai, Z., Bischof, C., Demmel, J., Dongarra, J., Croz, J. D., Greenbaum, A., Hammarling, S., McKenney, A., Ostrouchov, S., Sorensen, D., 1995. LAPACK Users’ Guide, 2nd Edition. SIAM, Philadelphia.
- Dhillon, I. S., 1997. A new  $o(n^2)$  algorithm for the symmetric tridiagonal eigenvalue/eigenvector problem. Ph.D. thesis, University of California, Berkeley, available from <http://www.cs.berkeley.edu/~inderjit/>.
- Fernando, K. V., 1997. On computing an eigenvector of a tridiagonal matrix. i. basic results. SIAM Journal on Matrix Analysis and Applications 18 (4), 1013–34.
- Fernando, K. V., 1998. Accurately counting singular values of bidiagonal matrices and eigenvalues of skew-symmetric tridiagonal matrices. SIAM Journal on Matrix Analysis and Applications 20 (2), 373–399.
- Godunov, S. K., Antonov, A. G., Kiriljuk, O. P., Kostin, V. I., 1988. Guaranteed Accuracy in the Solution of SLAE in Euclidean Spaces (In Russian). Nauka, Novosibirsk.
- Godunov, S. K., Antonov, A. G., Kiriljuk, O. P., Kostin, V. I., 1993. Guaranteed accuracy in numerical linear algebra. Kluwer Academic Publishers Group, Dordrecht, translated and revised from the 1988 Russian original.

- Golub, G. H., Loan, C. F. V., 1996. Matrix Computations, 3rd Edition. The Johns Hopkins University Press.
- Smith, B. T., Boyle, J. M., Dongarra, J. J., Garbow, B. S., Ikebe, Y., Klema, V. C., Moler, C. B., 1976. Matrix Eigensystem Routines – EISPACK Guide. Vol. 6 of Lecture Notes in Computer Science. Springer-Verlag, Berlin.
- Wilkinson, J. H., 1965. The Algebraic Eigenvalue Problem. Oxford University Press.