# Accuracy of Iterative Refinement of Eigenvalue Problems 

Jolanta Głuchowska-Jastrzębska, Alicja Smoktunowicz

## Streszczenie

We investigate numerical properties of Newton's algorithm for improving an eigenpair of a real matrix $A$ using only fixed precision arithmetic. We show that under natural assumptions it produces an eigenpair of a componentwise small relative perturbation of the data matrix $A$.

## 1 Introduction

We are interested in improving the accuracy of a given eigenvalue and its associated eigenvector. We consider the problem

$$
\begin{equation*}
A x=\lambda x, \quad x \neq 0 \tag{1}
\end{equation*}
$$

of evaluating a real eigenpair $\left(x^{*}, \lambda^{*}\right)$ for a real matrix $A \in \mathcal{R}^{n \times n}$ using only $t$-digit floating-point arithmetic $\mathrm{fl}(\mathrm{t})$. We assume that $\lambda^{*}$ is a single eigenvalue and $x^{*}$ is a corresponding eigenvector of the
matrix $A$.
Standard procedures produce such an eigenpair $(\tilde{x}, \tilde{\lambda})$ that

$$
\begin{equation*}
(A+\delta A) \tilde{x}=\tilde{x} \tilde{\lambda}, \quad\|\delta A\| \leq \rho K\|A\|, \tag{2}
\end{equation*}
$$

$K$ being a modest constant depending only on $n$, very often $K=\mathcal{O}\left(n^{2}\right)$ and $\rho=2^{-t}$ denotes the machine precision.

Throughout the paper we assume that || \| is a spectral vector-matrix norm. Note that the mere representation of the coefficients of the matrix $A$ in $\mathrm{fl}(\mathrm{t})$ arithmetic makes us solve

$$
\begin{equation*}
(A+E) x=\lambda x, \quad|E| \leq \rho|A|, \tag{3}
\end{equation*}
$$

i.e. $\left|e_{i, j}\right| \leq \rho\left|a_{i, j}\right|$ for $i, j=1, \ldots, n$, instead of the original problem (1). Here the elements of $E$ are small componentwise with respect to the coefficients of the matrix $A$. We ask whether it is possible to achieve the highest accuracy attainable in $\mathrm{fl}(\mathrm{t})$ which may be characterized as follows. The computed eigenpair $(\tilde{x}, \tilde{\lambda})$ of the matrix $A$ is the exact eigenpair of a slightly perturbed matrix $A+\Delta A$, i.e.

$$
\begin{equation*}
(A+\Delta A) \tilde{x}=\tilde{x} \tilde{\lambda}, \quad|\Delta A| \leq \rho K|A|, \tag{4}
\end{equation*}
$$

$K$ being at most of order $n^{2}$. We call such property componentwise backward stability. It is well- known [Cf. [9], [8]] that the property (2) is equivalent to the following relations

$$
\begin{equation*}
r=A \tilde{x}-\tilde{\lambda} \tilde{x}, \quad\|r\| \leq \rho\|A\|\|\tilde{x}\| K \tag{5}
\end{equation*}
$$

Since (2) is weaker than (4) we can conclude that it may occur that the normwise relative backward error

$$
\begin{equation*}
\frac{\|A \tilde{x}-\tilde{\lambda} \tilde{x}\|}{\rho\|A\|\|\tilde{x}\|} \tag{6}
\end{equation*}
$$

might be quite small, whereas the componentwise relative backward error

$$
\begin{equation*}
\max _{i} \frac{\left|(A \tilde{x}-\tilde{\lambda} \tilde{x})_{i}\right|}{\rho(|A \| \tilde{x}|)_{i}} \tag{7}
\end{equation*}
$$

is very large.

The idea of improving the computed solution is not quite new. The technique used is very similar to iterative refinement (IR) of solutions of linear systems $A x=b$ in fixed precision arithmetic. For brief survey of recent results see [2] and [8]. In fact, IR is a variant of Newton's method for solving the system $F(x)=0$, where $F(x)=A x-b$ (Cf. [12]). The Newton method is widely used for correction of computed solution. However, sometimes it is necessary to use higher precision (Cf. [1], [5]). A crucial point of the stability of Newton's method is how accurate the values of $F$ can be computed in $\mathrm{fl}(\mathrm{t})$ arithmetic.

It is known that also inverse iteration and the Rayleigh quotient iteration were introduced as ways of improving computed eigenpairs of matrices (Cf. [10], [11]). Several variants of Newton's method for the eigenvalue problem of a matrix were investigated (Cf. [3]-[5], [7], [11], [13]-[14]). However, only a little is known about numerical properties of the methods.

We consider the technique of iterative refinement proposed by Peters and Wilkinson (Cf. [11]). We stress that we use fixed precision arithmetic.

Note that the problem of evaluating an eigenpair is equivalent to this of solving the following nonlinear system $F(x, \lambda)=0$, where

$$
F(x, \lambda)=\binom{A x-\lambda x}{\left(1-x^{T} x\right) / 2}
$$

We apply Newton's method to the function $F$ and investigate the properties of the resulting algorithm. In Section 2 we define it in detail and consider its theoretical features. We show that the above algorithm provides the best solution to the problem (1) that can be obtained using merely the $t$-digit floating point arithmetic $\mathrm{fl}(\mathrm{t})$. Precisely, we prove that for the computed eigenpair ( $x_{k}, \lambda_{k}$ ) there exist perturbations $\Delta A_{k}, \Delta x_{k}$ such that

$$
\begin{equation*}
\limsup _{k \rightarrow \infty}\left\|\left(A+\Delta A_{k}\right)\left(x_{k}+\Delta x_{k}\right)-\lambda_{k}\left(x_{k}+\Delta x_{k}\right)\right\|=\mathcal{O}\left(\rho^{2}\right) \tag{8}
\end{equation*}
$$

and

$$
\begin{equation*}
\limsup _{k \rightarrow \infty}\left|\left(x_{k}+\Delta x_{k}\right)^{T}\left(x_{k}+\Delta x_{k}\right)-1\right|=\mathcal{O}\left(\rho^{2}\right) \tag{9}
\end{equation*}
$$

where $\left|\Delta A_{k}\right| \leq \rho L_{1}|A|,\left|\Delta x_{k}\right| \leq \rho L_{2}\left|x_{k}\right|$, and $L_{1}, \quad L_{2}$ are modest constants depending only on $n$.

Any algorithm evaluating an eigenpair $\left(x_{k}, \lambda_{k}\right)$ that satisfies the above property is called componentwise backward stable. It means that ignoring terms of order $\mathcal{O}\left(\rho^{2}\right)$, there exists $k^{*}$ such that for all $k \geq k^{*}$, the computed eigenpair ( $x_{k}, \lambda_{k}$ ) is an exact solution to a slightly perturbed eigenproblem. We show that the above algorithm, implemented with $t$-digit floating point arithmetic $\mathrm{fl}(\mathrm{t})$ is componentwise backward stable. However, the constant which appears in the " $\mathcal{O}$ " notation can be arbitrary, so we deal only with the asymptotic stability.

In Section 3 we show the way of obtaining of a version of the method which additional cost is only of order $\mathcal{O}\left(n^{2}\right)$ flops. This is the algorithm IREP (Iterative Refinement of Eigenvalue Problem).

## 2 Definition of the method and theoretical properties

We consider the problem (1) of evaluating a real eigenpair $(x, \lambda)$ of a real matrix $A \in \mathcal{R}^{n \times n}$. It is equivalent to this of solving the nonlinear system $F(x, \lambda)=0$ where

$$
\begin{equation*}
F(x, \lambda)=\binom{A x-\lambda x}{\left(1-x^{T} x\right) / 2} \tag{10}
\end{equation*}
$$

We apply Newton's method to the function $F$.
In this case the Jacobian matrix $F^{\prime}(x, \lambda)$ is of the form

$$
F^{\prime}(x, \lambda)=\left(\begin{array}{cc}
A-\lambda I & -x  \tag{11}\\
-x^{T} & 0
\end{array}\right)
$$

Note that if $A$ is symmetric then the Jacobian matrix $F^{\prime}(x, \lambda)$ is also symmetric.

Algorithm IREP.
We construct a sequence $\left(x_{k}, \lambda_{k}\right) \approx\left(x^{*}, \lambda^{*}\right)$ as follows.
(i) Start with: $\left(x_{0}, \lambda_{0}\right), x_{0}^{T} x_{0}=1$;

Set $k:=0$;
(ii) In the next step verify the termination criterion:

If for any element $\left|A x_{k}-\lambda_{k} x_{k}\right|>\rho 10 n\left|A \| x_{k}\right|$ then

- calculate

$$
F\left(x_{k}, \lambda_{k}\right)=\binom{A x_{k}-\lambda_{k} x_{k}}{\left(1-x_{k}^{T} x_{k}\right) / 2}
$$

- solve a linear system

$$
\left(\begin{array}{cc}
A-\lambda_{k} I & -x_{k} \\
-x_{k}^{T} & 0
\end{array}\right)\binom{\Delta x_{k}}{\Delta \lambda_{k}}=-F\left(x_{k}, \lambda_{k}\right)
$$

evaluating a correction $\left(\Delta x_{k}, \Delta \lambda_{k}\right)$;

- update the approximate

$$
\begin{aligned}
& x_{k+1}:=x_{k}+\Delta x_{k}, \\
& \lambda_{k+1}:=\lambda_{k}+\Delta \lambda_{k} ;
\end{aligned}
$$

(iii) Set $k:=k+1$ and return to (ii).

Convergence can be therefore analysed using the Newton-Kantorovitch Theorem (Cf. [1], [5], [14]). Such an analysis leads to a result of the following form: if the initial error is small enough then the iteration converges quadratically.
Now we estimate the condition number of the Jacobian matrix. If $\lambda^{*}$ is a simple eigenvalue of $A$ then the Jacobian matrix in (11) is nonsigular at ( $x^{*}, \lambda^{*}$ ), where $A x^{*}=$ $\lambda^{*} x^{*}$ and $\left\|x^{*}\right\|=1$ (Cf. [11], [14]).

Theorem 2.1 Let $y^{*}$ and $x^{*},\left\|x^{*}\right\|=\left\|y^{*}\right\|=1$, be respectively the left and right eigenvectors corresponding to the simple eigenvalue $\lambda^{*}$ of a real matrix $A \in \mathcal{R}^{n \times n}$.

Denote by $\sigma(A)=\left\{\lambda_{1}(A), \ldots, \lambda_{n}(A)\right\}$ the set of the eigenvalues of $A, \lambda^{*}=\lambda_{1}(A)$. Then the condition number cond $(B)=\|B\|\left\|B^{-1}\right\|$ of the Jacobian matrix

$$
\begin{equation*}
B=F^{\prime}\left(x^{*}, \lambda^{*}\right) \tag{12}
\end{equation*}
$$

is bounded by the condition number of $\lambda^{*}$ :

$$
\begin{equation*}
\operatorname{cond}(B) \geq \frac{\left\|y^{*}\right\|\left\|x^{*}\right\|}{\left|y^{* T} x^{*}\right|} \tag{13}
\end{equation*}
$$

Moreover, if $A$ is symmetric then

$$
\begin{equation*}
\operatorname{cond}(B)=\frac{\max _{i}\left\{1,\left|d_{i}\right|\right\}}{\min _{i}\left\{1,\left|d_{i}\right|\right\}} \tag{14}
\end{equation*}
$$

where $d_{i}=\lambda_{i}(A)-\lambda^{*}, i=2, \ldots, n$.
Proof. Since $\left(-x^{* T}, 0\right)^{T}$ is a column of $B$, so $\|B\| \geq\left\|\left(-x^{* T}, 0\right)\right\|=\left\|x^{*}\right\|$.
Let us define

$$
w^{T}=\left(y^{* T}, 0\right)
$$

Then

$$
w^{T} B=\left(0,-y^{* T} x^{*}\right),
$$

hence

$$
\left\|B^{-1}\right\| \geq \frac{\|w\|}{\left\|w^{T} B\right\|}=\frac{\left\|y^{*}\right\|}{\left|y^{* T} x^{*}\right|}
$$

From these we obtain the estimation (13).

Now we assume that $A$ is symmetric. Then there exists an orthogonal matrix $Q$ such that $A=Q D Q^{T}$ where $D=\operatorname{diag}\left(\lambda_{1}(A), \lambda_{2}(A), \ldots, \lambda_{n}(A)\right)$. Without loss of generality we can assume that $\left\|x^{*}\right\|=1$. From this and the assumption that $\lambda^{*}$ is a simple eigenvalue of $A$ (i.e. $A x^{*}=\lambda^{*} x^{*}$ and $\lambda^{*} \neq \lambda_{i}(A)$ for $i=2, \ldots, n$ ) it follows that $Q^{T} x^{*}= \pm e_{1}$.

Let us define the orthogonal matrix

$$
V=\left(\begin{array}{cc}
Q & 0 \\
0^{T} & 1
\end{array}\right)
$$

Then

$$
V^{T} B V==\left(\begin{array}{cc}
D-\lambda^{*} I & \pm e_{1} \\
\pm e_{1}^{T} & 0,
\end{array}\right)
$$

so the eigenvalues of $B$ are equal to $\pm 1$ and $\lambda_{i}(A)-\lambda^{*}, i=2, \ldots, n$. This immediately proves (14).

## Remark.

If $A$ is symmetric then we can write $B^{-1}$ as follows:

$$
B^{-1}=\left(\begin{array}{cc}
\left(A-\lambda^{*} I\right)^{+} & -x^{*}  \tag{15}\\
-x^{* T} & 0
\end{array}\right) .
$$

## 3 Implementation of algorithm IREP

Algorithm IREP may be implemented in several ways. At every stage of IREP we have to solve a linear system of equations $J z=f$ where $J$ is the Jacobian matrix:

$$
J=\left(\begin{array}{cc}
A-\lambda I & -x \\
-x^{T} & 0
\end{array}\right) .
$$

Now we propose an efficient algorithm (S) which is normwise backward stable.

Algorithm S of solving a linear system $J z=f$.
(a) Reduction of the matrix $A$ to Hessenberg for by Householder algorithm with $\operatorname{cost} \mathcal{O}\left(n^{3}\right)$ flops. We obtain $A=Q H Q^{T}$ where $Q^{T} Q=I$ and $H$ being upper Hessenberg matrix.
(b) Let us define

$$
V=\left(\begin{array}{cc}
Q & 0  \tag{16}\\
0^{T} & 1
\end{array}\right), C=V^{T} J V, y=-Q^{T} x \text {. }
$$

Then

$$
C=\left(\begin{array}{cc}
H-\lambda I & y  \tag{17}\\
y^{T} & 0
\end{array}\right)
$$

Thus the algorithm for solving $J z=f$ is the following:

- Compute $g=V^{T} f$;
if $f=\left(u^{T}, f_{n+1}\right)^{T}$ then $g=\left(u^{T} Q, f_{n+1}\right)^{T}$.
- Evaluate $w$ solving a linear system $C w=g$.

The matrix $C$ is namely of the form

$$
C=\left(\begin{array}{cccccc}
c_{1,1} & c_{1,2} & c_{1,3} & \ldots & c_{1, n} & y_{1} \\
c_{2,1} & c_{2,2} & c_{2,3} & \ldots & c_{2, n} & y_{2} \\
0 & c_{3,2} & c_{3,3} & \ldots & c_{3, n} & y_{3} \\
\ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\
0 & 0 & 0 & c_{n, n-1} & c_{n, n} & y_{n} \\
y_{1} & y_{2} & y_{3} & \ldots & y_{n} & 0
\end{array}\right) .
$$

We permute the rows of the matrix $C$ (and, respectively, of the right hand side vector) in such a way that the obtained matrix $\hat{C}$ has upper Hessenberg form with additional co-diagonal. Namely, the matrix $\hat{C}$ is of the form

$$
\hat{C}=\left(\begin{array}{cccccc}
y_{1} & y_{2} & y_{3} & \ldots & y_{n} & 0 \\
c_{1,1} & c_{1,2} & c_{1,3} & \ldots & c_{1, n} & y_{1} \\
c_{2,1} & c_{2,2} & c_{2,3} & \ldots & c_{2, n} & y_{2} \\
\ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\
\ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\
0 & 0 & 0 & c_{n, n-1} & c_{n, n} & y_{n}
\end{array}\right) .
$$

We solve the system under consideration with $\mathcal{O}\left(n^{2}\right)$ flops using GEPP (Gaussian elimination with partial pivoting).

- Find $z=V w$;

$$
\text { if } w=\left(c^{T}, w_{n+1}\right)^{T} \text { then } z=\left(c^{T} Q, w_{n+1}\right)^{T} .
$$

## 4 Numerical properties of Newton's method

We specify the conditions under which the algorithm IREP is componentwise backward stable. This is the topic of Theorem 4.2 dealing with numerical properties of Newton's method. The problem is considered as this of evaluating a simple root $z^{*}$ of the
nonlinear system $F(z)=0$, where $F: \mathbf{R}^{m} \rightarrow \mathbf{R}^{m}$. We assume that the function $F$ depends parametrically on a data vector $d$, i.e. $F(z)=F(z ; d)$, where $d \in D \subset \mathbf{R}^{s}$ and is sufficiently smooth in $d$ and $z$. It is sufficient to assume that $F$ has a first Lipschitz derivative in a neighbourhood of such a point $\left(z^{*} ; d\right)$ for which $F\left(z^{*} ; d\right)=0$. In our problem this assumption is fulfilled (with Lipschitz constant equal to $\sqrt{2}$ ). It is natural to assume that data are all coefficients $a_{i, j}$ of the matrix $A$. More exactly, we put $d=\left(a_{1,1}, \ldots, a_{1, n}, a_{2,1}, \ldots, a_{n, n}\right)^{T}$.

For the detailed discussion on the data vector choice and on numerical properties of Newton's iteration in normwise sense see [16].

Recall Newton's method for the function $F$ (Cf. [1]], [16]):

- compute $F\left(z_{k}\right)$ and $F^{\prime}\left(z_{k}\right)$,
- evaluate the correction $\Delta_{k}$ :

$$
F^{\prime}\left(z_{k}\right) \Delta_{k}=-F\left(z_{k}\right)
$$

- update the approximate: $z_{k+1}=z_{k}+\Delta_{k}$.

Here and throughout the paper we denote by $f l(F(z ; d))$ a value of $F(z ; d)$ which is computed in $t$-digit floating point binary arithmetic.

The following Theorem 4.2 can be stated as a modification of Woźniakowski's theorem (Cf. [16]).

Theorem 4.1 Let

$$
\begin{equation*}
f l\left(F\left(z_{k}\right)\right)=f l\left(F\left(z_{k} ; d\right)\right)=\left(I+\delta F_{k}\right) F\left(z_{k}+\delta z_{k} ; d+\delta d_{k}\right), \tag{18}
\end{equation*}
$$

where

$$
\begin{equation*}
\left|\delta F_{k}\right| \leq \rho L_{1} I,\left|\delta z_{k}\right| \leq \rho L_{2}\left|z_{k}\right|,\left|\delta d_{k}\right| \leq \rho L_{3}|d|, \tag{19}
\end{equation*}
$$

$L_{1}, L_{2}, L_{3}$ being some constants dependent on the problem dimension and independent of $k, z_{k}, d$ and $\rho$.

Assume that in every iteration the Jacobian matrix is evaluated in such a way that

$$
\begin{equation*}
f l\left(F^{\prime}\left(z_{k} ; d\right)=F^{\prime}\left(z_{k} ; d\right)+\delta F_{k}^{\prime},\left\|\delta F_{k}^{\prime}\right\|=\mathcal{O}(\rho) .\right. \tag{20}
\end{equation*}
$$

At last, let the algorithm for evaluating the correction $\Delta_{k}$ satisfies

$$
\begin{equation*}
\left(F^{\prime}\left(z_{k}\right)+\Delta F_{k}^{\prime}\right) \Delta_{k}=-f l\left(F\left(z_{k}\right)\right),\left\|\Delta F_{k}^{\prime}\right\|=\mathcal{O}(\rho) \tag{21}
\end{equation*}
$$

and the next approximate $z_{k+1}$ be computed as follows:

$$
\begin{equation*}
z_{k+1}=\left(I+D_{k}\right)\left(z_{k}+\Delta_{k}\right),\left|D_{k}\right| \leq \rho I \tag{22}
\end{equation*}
$$

Then the Newton algorithm of evaluating a simple zero $\alpha$ is componentwise stable, i.e. there exist such perturbations $\left\{z_{k}\right\}$ and $\left\{\Delta d_{k}\right\}$ that

$$
\begin{equation*}
\limsup _{k \rightarrow \infty} \| F\left(z_{k}+\Delta z_{k} ; d+\Delta d_{k} \|=\mathcal{O}\left(\rho^{2}\right)\right. \tag{23}
\end{equation*}
$$

where for sufficiently large $k$ :

$$
\begin{equation*}
\left|\Delta z_{k}\right| \leq \rho\left(L_{2}+1\right)\left|z_{k}\right|, \quad\left|\Delta d_{k}\right| \leq \rho L_{3}|d| \tag{24}
\end{equation*}
$$

The above result shows that for $k$ large enough a slightly perturbed computed value is the exact one of a slightly perturbed inputs (all perturbations are small componentwise).
Proof. We prove that "one step is enough" for Newton's method.
Assume that $z_{k}$ is close enough to a simple zero $\alpha$ of a "smooth" function $F$, so $\left\|z_{k}-\alpha\right\|=\mathcal{O}(\rho)$.

We want to find perturbations $\Delta z_{k+1}$ and $\Delta d_{k+1}$ such that

$$
\begin{equation*}
\left\|F\left(z_{k+1}+\Delta z_{k+1} ; d+\Delta d_{k+1}\right)\right\|=\mathcal{O}\left(\rho^{2}\right) \tag{25}
\end{equation*}
$$

where

$$
\begin{equation*}
\left|\Delta z_{k+1}\right| \leq \rho\left(L_{2}+1\right)\left|z_{k+1}\right|,\left|\Delta d_{k+1}\right| \leq \rho L_{3}|d| \tag{26}
\end{equation*}
$$

Let us define

$$
\hat{z}_{k+1}=\left(I+D_{k}\right)^{-1} z_{k+1} .
$$

Then

$$
F\left(\hat{z}_{k+1}+\delta z_{k} ; d+\delta d_{k}\right)=F\left(z_{k}+\Delta_{k}+\delta z_{k} ; d+\delta d_{k}\right)
$$

so using the Taylor extension we obtain

$$
F\left(\hat{z}_{k+1}+\delta z_{k} ; d+\delta d_{k}\right)=F\left(z_{k}\right)+F^{\prime}\left(z_{k}\right)\left(\Delta_{k}+\delta z_{k}\right)+F_{d}^{\prime}\left(z_{k} ; d\right) \delta d_{k}+\mathcal{O}\left(\rho^{2}\right)
$$

Notice that $F^{\prime}\left(z_{k}\right) \Delta_{k}=-f l\left(F\left(z_{k} ; d\right)\right)+\mathcal{O}\left(\rho^{2}\right)$ and $F\left(z_{k}\right)=\mathcal{O}(\rho)$, hence from (18) we have

$$
f l\left(F\left(z_{k} ; d\right)\right)=F\left(z_{k}\right)+\delta F_{k} F\left(z_{k}\right)+F^{\prime}\left(z_{k}\right) \delta z_{k}+F_{d}^{\prime}\left(z_{k} ; d\right) \delta d_{k}+\mathcal{O}\left(\rho^{2}\right)
$$

This equation may be written as follows

$$
f l\left(F\left(z_{k} ; d\right)\right)=F\left(z_{k}\right)+F^{\prime}\left(z_{k}\right) \delta z_{k}+F_{d}^{\prime}\left(z_{k} ; d\right) \delta d_{k}+\mathcal{O}\left(\rho^{2}\right) .
$$

From this is follows that

$$
F\left(\hat{z}_{k+1}+\delta z_{k} ; d+\delta d_{k}\right)=\mathcal{O}\left(\rho^{2}\right)
$$

Let us define

$$
\Delta d_{k+1}=\delta d_{k}
$$

and

$$
\hat{z}_{k+1}+\delta z_{k}=z_{k+1}+\mu_{k+1} .
$$

We can split $\mu_{k+1}$ as follows

$$
\mu_{k+1}=\Delta z_{k+1}+\mathcal{O}\left(\rho^{2}\right)
$$

where

$$
\left|\Delta z_{k+1}\right| \leq \rho\left(1+L_{2}\right)\left|z_{k+1}\right| .
$$

Then we have (25)-(26) which completes the proof.

## 5 Componentwise backward stability of IREP

We show that under some natural assumptions IREP is componentwise stable. We consider numerical implementation of Newton's method for $F$ defined by (10).

Note that we can evaluate $F^{\prime}(z)$ in such a way that a value $f l\left(F^{\prime}(z)\right)$ computed in $f l(t)$ satisfies

$$
f l\left(F^{\prime}(z)\right)=F^{\prime}(z)+\delta F^{\prime},\left|\delta F^{\prime}\right| \leq \rho\left|F^{\prime}(z)\right| .
$$

In particular, for the method under consideration the above theorem takes the following form.

Theorem 5.1 Suppose that the algorithm (solver $\mathbf{S}$ ) evaluates the correction $\Delta_{k}$ in such a way that

$$
\begin{equation*}
\left(F^{\prime}\left(z_{k}\right)+\Delta F_{k}^{\prime}\right) \Delta_{k}=-f l\left(F\left(z_{k}\right)\right),\left\|\Delta F_{k}^{\prime}\right\|=\mathcal{O}(\rho) . \tag{27}
\end{equation*}
$$

Assume that evaluating $F\left(z_{k}\right.$ we multiply the matrix $A$ by the vector $x$ in such a way that

$$
\begin{equation*}
f l(A x)=(A+E) x \quad|E| \leq \rho L|A|, \tag{28}
\end{equation*}
$$

where $L=$ const.
Then IREP using only $f(t)$ arithmetic produces a sequence $\left\{z_{k}\right\}$ such that for each computed eigenpair $\left(x_{k}, \lambda_{k}\right)$ there exist perturbations $\Delta A_{k}, \Delta x_{k}$ such that

$$
\begin{equation*}
\limsup _{k \rightarrow \infty}\left\|\left(A+\Delta A_{k}\right)\left(x_{k}+\Delta x_{k}\right)-\lambda_{k}\left(x_{k}+\Delta x_{k}\right)\right\|=\mathcal{O}\left(\rho^{2}\right) \tag{29}
\end{equation*}
$$

and

$$
\begin{equation*}
\limsup _{k \rightarrow \infty}\left|\left(x_{k}+\Delta x_{k}\right)^{T}\left(x_{k}+\Delta x_{k}\right)-1\right|=\mathcal{O}\left(\rho^{2}\right), \tag{30}
\end{equation*}
$$

where

$$
\begin{equation*}
\left|\Delta A_{k}\right| \leq \rho(L+2)|A|,\left|\Delta x_{k}\right| \leq \rho\left(1+\frac{n}{2}\right)\left|x_{k}\right| . \tag{31}
\end{equation*}
$$

Proof. For the sake of simplicity we omit the step index $k$ putting $x=x_{k}$ and $\lambda=\lambda_{k}$.

It is sufficient to show that using $f l(t)$ arithmetic we can compute the value of the function $F(z)$ in such a way that

$$
\begin{equation*}
f l(F(z))=(I+\delta F)\binom{(A+\delta A)(x+\delta x)-\lambda(x+\delta x)}{\left(1-(x+\delta x)^{T}(x+\delta x)\right) / 2} \tag{32}
\end{equation*}
$$

where

$$
\begin{equation*}
|\delta F| \leq \rho\left(2+\frac{n}{2}\right) I,|\delta A| \leq \rho(L+1)|A|,|\delta x| \leq \rho \frac{n}{2}|x| . \tag{33}
\end{equation*}
$$

Using common algorithm for summation of $n$ numbers we have (Cf. [8], [16]):

$$
f l\left(x^{T} x\right)=\left(x^{T} x\right)(1+\mu),|\mu| \leq \rho n .
$$

We also have

$$
f l(\lambda x)=\left(I+D_{1}\right) \lambda x,\left|D_{1}\right| \leq \rho I .
$$

From this it follows

$$
f l(F(z))=\binom{\left(I+D_{2}\right)\left\{(A+E) x-\left(I+D_{1}\right) \lambda x\right\}}{\left(1-x^{T} x(1+\mu)\right)(1+\epsilon) / 2}
$$

where $\left|D_{2}\right| \leq \rho I,|\epsilon| \leq \rho$.

Let us define

$$
\begin{gathered}
x+\delta x=x \sqrt{1+\mu}, \quad A+\delta A=\left(I+D_{1}\right)^{-1}(A+E), \\
I+D=\left(1+D_{2}\right)\left(1+D_{1}\right) / \sqrt{1+\mu},
\end{gathered}
$$

and

$$
I+\delta F=\left(\begin{array}{cc}
I+D & 0 \\
0 & 1+\epsilon
\end{array}\right)
$$

This proves the theorem.

## 6 Concluding remarks

In this paper, we have presented an algorithm IREP for improving an eigenpair of a real matrix using fixed aritmetic.

Numerical experiments that were carried out in MATLAB ( $\rho=2.2 e-16$ ) confirm theoretical considerations. We observed that the poor performance appeared only as a result of the closeness of eigenvalues of $A$ and in a case when the componentwise relative backward error was greater then $1.0 e 14$. The eigenvalues of $A$ were computed using a function eig.m in MATLAB. The eigenvalues, one at a time, were then been improved by the iterative refinement algorithm IREP. Very often, one or two steps were sufficient to terminate successfully the process.

## REFERENCES

[ 1 ] J. L. Barlow, Error analysis of update methods for the symmetric eigenvalue problem, SIAM J. Matrix Anal. Appl., Vol. 14, No. 2 (1993), 598-618.
[ 2 ] À. Björck, Iterative refinement and reliable computing, M. G. Cox and S. J. Hammarling, eds.., Oxford Univ. Press (1990), pp. 249-266.
[ 3 ] F. Chatelin, Simultaneous Newton's iteration for the eigenproblem, Computing, Suppl., 5 (1984), pp. 67-74.
[ 4 ] J. W. Demmel, Three methods for refining estimates of invariant subspaces, Computing, 38 (1987), pp. 43-57.
[5] J. J. Dongarra, C. B. Moler, J. H. Wilkinson, Improving the accuracy of computed eigenvalues and eigenvectors, SIAM J. Num. Anal., 20 (1983), pp. 46-58.
[ 6 ] J. Głuchowska, Numerical properties of Newton's algorithm for improving an eigenpair of a real matrix, Warsaw University, Manuscript, 1989.
[ 7 ] W. W. Hager, Applied Numerical Linear Algebra, Prentice-Hall, Englewood Cliffs, NJ, 1988.
[ 8 ] N. J. Higham, Accuracy and Stability of Numerical Algorithms, Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 1996.
[ 9 ] W. Oettli and W. Prager, Compatibility of approximate solutions of linear equations with given error bounds for coefficients and right-hand sides, Numer.Math., 6 (1964), pp. 405-409.
[10] B. Parlett, The symmetric eigenvalue problem, Prentice-Hall, Englewood Cliffs, NJ, 1980.
[ 11 ] G. Peters, J. H. Wilkinson, Inverse iteration, ill-conditioned equations and Newton's method, SIAM Rev., 21 (1979), pp. 339-360.
[12] A. Smoktunowicz and J. Sokolnicka, Binary cascades iterative refinement in doubled-mantissa arithmetics, BIT (1984), pp. 123-127.
[13] R. A. Tapia and D. L. Whitley, The projected Newton method has order $1+\sqrt{2}$ for the symmetric eigenvalue problem, SIAM J. Numer. Anal., 25 (1988), pp. 1376-1382.
[14] T. Yamamoto, Error bounds for computed eigenvalues and eigenvectors, Numer. Math. 34 (1980), pp. 189-199.
[ 15 ] J. H. Wilkinson, The algebraic eigenvalue problem, Oxford, Clarendon Press, 1965.
[ 16 ] H. Woźniakowski, Numerical stability for solving nonlinear equations, Numer. Math. 27 (1977), pp. 373-390.

Agricultural University of Warsaw, Wydział Ekonomiczno-Rolniczy, Katedra Ekonometrii i Informatyki, Nowoursynowska 166, Warsaw, Poland Institute of Mathematics,

Warsaw University of Technology, Pl.Politechniki 1, 00-661 Warsaw, Poland
(smok@im.pw.edu.pl)

