Newton's Method in Floating Point Arithmetic and Iterative Refinement of Generalized Eigenvalue Problems

Tisseur, Françoise Tisseur

2001

MIMS EPrint: **2007.223**

Manchester Institute for Mathematical Sciences
School of Mathematics

The University of Manchester

Reports available from: http://eprints.maths.manchester.ac.uk/
And by contacting: The MIMS Secretary
School of Mathematics
The University of Manchester
Manchester, M13 9PL, UK

ISSN 1749-9097

SIAM J. MATRIX ANAL. APPL.
Vol. 22, No. 4, pp. 1038–1057

NEWTON'S METHOD IN FLOATING POINT ARITHMETIC AND ITERATIVE REFINEMENT OF GENERALIZED EIGENVALUE PROBLEMS*

FRANÇOISE TISSEUR†

Abstract. We examine the behavior of Newton's method in floating point arithmetic, allowing for extended precision in computation of the residual, inaccurate evaluation of the Jacobian and unstable solution of the linear systems. We bound the limiting accuracy and the smallest norm of the residual. The application that motivates this work is iterative refinement for the generalized eigenvalue problem. We show that iterative refinement by Newton's method can be used to improve the forward and backward errors of computed eigenpairs.

Key words. Newton's method, generalized eigenvalue problem, iterative refinement, Cholesky method, backward error, forward error, rounding error analysis, limiting accuracy, limiting residual

AMS subject classifications. 65F15, 65F35

PII. S0895479899359837

1. Introduction.

This work is motivated by the symmetric definite generalized eigenvalue problem $Ax = \lambda Bx$ (A and B symmetric and one of them positive definite), for which no method is known that takes advantage of the symmetry, is efficient, and is backward stable. For the special case where both matrices are positive definite, such a method is available [26]. The aim is to show that iterative refinement by Newton's method can be used to improve the forward and backward errors of computed eigenpairs. An important question is how accurately the residuals must be evaluated in order to improve the relative forward error and/or the backward error.

For added generality we give a detailed analysis of the general Newton method in floating point arithmetic, allowing for extended precision in computation of the residual, possibly inaccurate evaluation of the Jacobian and unstable linear system solvers. We bound the limiting accuracy that can be obtained and the smallest norm of the residual.

Lancaster [19], Woźniakowski [28], Ypma [29], [30], and Dennis and Walker [6] have also considered the effects of inaccuracy, computational or otherwise, on Newton's method for solving nonlinear algebraic equations. None of these authors analyzed the behavior of the residual. Lancaster and Ypma were interested in how the approximate iterate is related to the exact one rather than the error in the approximate iterate. Woźniakowski carried out his analysis with the big-Oh notation and therefore his results contain unknown constants. We follow the same approach as Dennis and Walker [6] in that our results are based directly on the error in the computed iterates. The analysis in [6] is very general and uses several assumptions and constants that are difficult to interpret and understand even for the special case discussed therein (iterative refinement for linear systems of equations).

The residual contains information that is crucial for improving an approximate solution by Newton's method. Thus it should be computed as accurately as possible.

*Received by the editors August 4, 1999; accepted for publication (in revised form) by J. Varah October 5, 2000; published electronically February 23, 2001.
http://www.siam.org/journals/simax/22-4/35983.html

†Department of Mathematics, University of Manchester, Manchester, M13 9PL, England (ftisseur@ma.man.ac.uk, http://www.ma.man.ac.uk/~ftisseur/). This work was supported by Engineering and Physical Sciences Research Council grant GR/L76532.

Recently, mixed precision BLAS (XBLAS) routines have been proposed as a standard [2], where extended precision arithmetic is used internally to the BLAS and then the output is rounded to working precision. These new BLAS make the computation of the residual in mixed precision feasible for many problems, including the generalized eigenvalue problem considered here.

We first rework the forward error analysis of [6] for Newton's method in floating point arithmetic. We use different assumptions that are more appropriate when we have access to extended precision in computation of the residual and when we are using a possibly unstable linear system solver. The results we obtain are of more practical use than those in [6], [19], [28], [29] but consistent with them. We also estimate the limiting accuracy that can be obtained near a solution.

Next, we study the convergence of the norm of the residual, bounding the smallest norm. For many problems the backward error is a scaled residual norm, in which case we can use our results to bound the backward error. The idea of using iterative refinement to obtain a small backward error with a potentially unstable solution method has been investigated for linear systems by several authors, including Jankowski and Woźniakowski [18], Skeel [22], and Higham [17], and more recently for the algebraic Riccati equation by Ghavimi and Laub [11]. The idea does not seem to have been applied previously to the generalized eigenvalue problem.

In section 3 we apply our results to linear systems and to the standard and generalized eigenvalue problems. In section 4 we present numerical examples for the symmetric definite eigenvalue problem that motivated the whole analysis.

2. Newton's method in floating point arithmetic.

2.1. Basics and notation. We begin by describing our notation. Let $F : \mathbb{R}^m \mapsto \mathbb{R}^m$ be continuously differentiable on $\mathbb{R}^m$. We denote by J the Jacobian matrix $(\partial F_i/\partial v_j)$ of F and assume that J is Lipschitz continuous with constant β in $\mathbb{R}^m$, that is,

$$\|J(w) - J(v)\| \le \beta\|w - v\| \quad \text{for all} \quad v, w \in \mathbb{R}^m,$$

where $\|\cdot\|$ denotes any vector norm and the corresponding operator norm. We denote by $\kappa(J) = \|J\|\|J^{-1}\|$ the condition number of the matrix J. We attempt to solve the system of nonlinear equations $F(v) = 0$ by Newton's method:

$$J(v_i)(v_{i+1} - v_i) = -F(v_i), \quad i \ge 0, \tag{2.1}$$

where v_0 is given. We implement (2.1) as

$$\begin{aligned} &\text{Solve } J(v_i)d_i = -F(v_i),\\ &v_{i+1} = v_i + d_i. \end{aligned}$$

Newton's method is attractive because under appropriate conditions it converges rapidly from any sufficiently good initial guess. In particular, if the Jacobian is nonsingular at the solution, local quadratic convergence can be proved [5, Thm. 5.2.1]. The Kantorovich theorem yields a weaker bound on the convergence rate but makes no assumption on the nonsingularity of Jacobian at the solution [5, Thm. 5.3.1], [24].

We use hats to denote computed quantities. We work with the standard model of floating point arithmetic [16, section 2.3]

$$fl(x \operatorname{op} y) = (x \operatorname{op} y)(1 + \delta), \qquad |\delta| \le u, \quad \operatorname{op} = +, -, *, /,$$

where u is the unit roundoff.

In floating point arithmetic, we have

$$\widehat{v}_{i+1} = \widehat{v}_i - (J(\widehat{v}_i) + E_i)^{-1}\,(F(\widehat{v}_i) + e_i) + \varepsilon_i, \tag{2.2}$$

where

- e_i is the error made when computing the residual $F(\widehat{v}_i)$,
- E_i is the error incurred in forming $J(\widehat{v}_i)$ and solving the linear system for d_i,
- ε_i is the error made when adding the correction $\widehat{d}_i$ to $\widehat{v}_i$.

We assume that $F(\widehat{v}_i)$ is computed in the possibly extended precision $\bar{u} \le u$ before rounding back to working precision u, and that $\widehat{d}_i, \widehat{v}_i$ are computed at precision u. Hence we assume that there exists a function ψ depending on $F, \widehat{v}_i, u$, and $\bar{u}$ such that

$$\|e_i\| \le u\|F(\widehat{v}_i)\| + \psi(F, \widehat{v}_i, u, \bar{u}). \tag{2.3}$$

Note that standard error analysis shows that $\|e_i\| \le u\|F(\widehat{v}_i)\|$ is the best we can obtain in practice for both mixed and fixed precision. Later, we will give an explicit formula for ψ in the case of linear systems and the generalized eigenvalue problem. We assume that the error E_i satisfies

$$\|E_i\| \le u\phi(F, \widehat{v}_i, n, u) \tag{2.4}$$

for some function ϕ that reflects both the instability of the linear solver and the error made when approximating or forming $J(\widehat{v}_i)$. In practice, we certainly have $\phi(F, \widehat{v}_i, n, u) \ge \|J(\widehat{v}_i)\|$. For the error ε_i we have

$$\|\varepsilon_i\| \le u(\|\widehat{v}_i\| + \|\widehat{d}_i\|).$$

We will make use of the constants

$$\gamma_n = \frac{cnu}{1 - cnu} \quad \text{and} \quad \bar{\gamma}_n = \frac{cn\bar{u}}{1 - cn\bar{u}}, \tag{2.5}$$

where c is a small integer constant.

2.2. Forward error. First we consider the change in error for a single step of an iteration of the form (2.2). For notational convenience we write $v = \widehat{v}_i$, $\bar{v} = \widehat{v}_{i+1}$, and

$$\bar{v} = v - (J + E)^{-1}(r + e) + \varepsilon, \tag{2.6}$$

where $r = F(v), J = J(v)$, and

$$\|E\| \le u\phi(F, v, n, u), \tag{2.7}$$

$$\|e\| \le u\|r\| + \psi(F, v, u, \bar{u}), \quad \|\varepsilon\| \le u(\|v\| + \|d\|),$$

with

$$d = (J + E)^{-1}(r + e). \tag{2.8}$$

We will often refer to the following lemma.

LEMMA 2.1 (see [5, Lem. 4.1.12]). *For any $v, w \in \mathbb{R}^m$,*

$$\|F(w) - F(v) - J(v)(w - v)\| \le \frac{\beta}{2}\|w - v\|^2. \tag{2.9}$$

THEOREM 2.2. *Assume that there is a v_* such that $F(v_*) = 0$, $J_* = J(v_*)$ is nonsingular, and*

$$\|J^{-1}E\| \le \nu < 1. \tag{2.10}$$

Then, for all v such that

$$\beta\|J_*^{-1}\|\|v - v_*\| \le \mu < 1, \tag{2.11}$$

$\bar{v}$ in (2.6) is well defined and

$$\|\bar{v} - v_*\| \le G\|v - v_*\| + g,$$

where

$$G = \frac{1}{1-\nu}\|J^{-1}E\| + \frac{(1+u)^2}{2(1-\mu)(1-\nu)}\beta\|J_*^{-1}\|\|v - v_*\| + \frac{u(2+u)}{(1-\mu)(1-\nu)}\kappa(J_*) + u$$

and

$$g = \frac{1+u}{(1-\mu)(1-\nu)}\|J_*^{-1}\|\psi(F, v, u, \bar{u}) + u\|v_*\|.$$

Proof. From assumption (2.11) and the Lipschitz property of J we have

$$\|J_*^{-1}(J - J_*)\| \le \beta\|J_*^{-1}\|\|v - v_*\| \le \mu < 1. \tag{2.12}$$

From the identity

$$J = J_*(I + J_*^{-1}(J - J_*)) \tag{2.13}$$

it then follows that J is nonsingular with inverse given by

$$J^{-1} = (I + J_*^{-1}(J - J_*))^{-1}J_*^{-1}$$

and with

$$\|J^{-1}\| \le \frac{\|J_*^{-1}\|}{1 - \|J_*^{-1}(J - J_*)\|} \le \frac{1}{1-\mu}\|J_*^{-1}\|. \tag{2.14}$$

Similarly, assumption (2.10) guarantees that $J + E$ is nonsingular and that, using (2.14),

$$\|(J+E)^{-1}\| \le \frac{\|J^{-1}\|}{1 - \|J^{-1}E\|} \le \frac{1}{(1-\mu)(1-\nu)}\|J_*^{-1}\|. \tag{2.15}$$

Since $(J+E)^{-1}$ exists, $\bar{v}$ in (2.6) is well defined. We have

$$\begin{aligned}\bar{v} - v_* &= v - v_* - (J+E)^{-1}(r+e) + \varepsilon \\ &= (I - (J+E)^{-1}J)(v - v_*) - (J+E)^{-1}(r - J(v - v_*) + e) + \varepsilon,\end{aligned}$$

which gives

$$\|\bar{v} - v_*\| \le \|I - (J+E)^{-1}J\|\|v - v_*\| + \|(J+E)^{-1}\|(\|r - J(v - v_*)\| + \|e\|) + \|\varepsilon\|.$$

From

$$I - (J+E)^{-1}J = (J+E)^{-1}E = (I + J^{-1}E)^{-1}J^{-1}E$$

it follows that

$$\|I - (J+E)^{-1}J\| \le \frac{1}{1-\nu}\|J^{-1}E\|.$$

From Lemma 2.1,

$$\|r - J(v-v_*)\| \le \frac{\beta}{2}\|v-v_*\|^2 \quad \text{and} \quad \|r - J_*(v-v_*)\| \le \frac{\beta}{2}\|v-v_*\|^2,$$

so that

$$\|r\| \le \|r - J_*(v-v_*)\| + \|J_*(v-v_*)\| \le \frac{\beta}{2}\|v-v_*\|^2 + \|J_*\|\|v-v_*\| \tag{2.16}$$

and hence

$$\|e\| \le u\left(\frac{\beta}{2}\|v-v_*\|^2 + \|J_*\|\|v-v_*\|\right) + \psi(F,v,u,\bar{u}).$$

We have

$$\|\varepsilon\| \le u(\|v-v_*\| + \|v_*\| + \|d\|)$$

with

$$\begin{aligned}\|d\| &\le \|(J+E)^{-1}\|(\|r\| + \|e\|) \\ &\le \|(J+E)^{-1}\|\big((1+u)\|r\| + \psi(F,v,u,\bar{u})\big) \\ &\le \frac{1}{(1-\mu)(1-\nu)}\|J_*^{-1}\|\left[(1+u)\left(\frac{\beta}{2}\|v-v_*\| + \|J_*\|\right)\|v-v_*\| \right. \\ &\qquad\qquad \left. + \psi(F,v,u,\bar{u})\right],\end{aligned} \tag{2.17}$$

using (2.15) and (2.16). Hence,

$$\|\bar{v} - v_*\| \le G\|v-v_*\| + g,$$

where G and g are given in the statement of the theorem. ∎

Assumptions (2.10) and (2.11) are necessary for $\bar{v}$ in (2.6) to be defined. Assumption (2.10) is a condition on the stability of the linear system solver and the accuracy of the Jacobian.

In exact arithmetic we have $u = \psi(F,v,u,\bar{u}) = \nu = 0$ and $E = 0$. Then, for $\mu \le 1/2$, Theorem 2.2 reduces to the local quadratic convergence theorem for Newton's method [5, Thm. 5.2.1] applied to a single step.

Clearly, for $\mu \le \frac{1}{8}$, $\nu \le \frac{1}{8}$, if J_* is not too ill conditioned, say, $u\kappa(J_*) \le \frac{1}{8}$, then we have $G \le \frac{1}{2}$. Thus the error contracts unless $g \gtrsim \|v - v_*\|$. Hence, the best limiting normwise accuracy we can guarantee is

$$\frac{g}{\|v_*\|} = \frac{1+u}{(1-\mu)(1-\nu)}\frac{\|J_*^{-1}\|}{\|v_*\|}\psi(F,v,u,\bar{u}) + u,$$

which depends on the accuracy with which the residual is computed. If $\|J_*^{-1}\|\psi(F, v, u, \bar{u}) \leq cu\|v_*\|$ for some constant c, then we can expect to obtain a normwise relative error of order cu.

Note that the rate of convergence depends on the accuracy of the Jacobian and on the stability of the linear system solver, since G depends strongly on E, but the limiting accuracy is essentially independent of the solver (for $\nu < \frac{1}{8}$, say). Note also that G is independent of $\bar{u}$, which means that the rate of convergence is bounded independent of the precision used to compute the residual.

COROLLARY 2.3. *Assume that there is a v_* such that $F(v_*) = 0$ and $J_* = J(v_*)$ is nonsingular and satisfies*

$$u\kappa(J_*) \leq \frac{1}{8}. \tag{2.18}$$

Assume also that for ϕ in (2.4),

$$u\|J(\widehat{v}_i)^{-1}\|\phi(F, \widehat{v}_i, n, u) \leq \frac{1}{8} \quad \textit{for all } i. \tag{2.19}$$

Then, for all v_0 such that

$$\beta\|J_*^{-1}\|\|v_0 - v_*\| \leq \frac{1}{8}, \tag{2.20}$$

Newton's method in floating point arithmetic generates a sequence $\{\widehat{v}_{i+1}\}$ whose normwise relative error decreases until the first i for which

$$\frac{\|\widehat{v}_{i+1} - v_*\|}{\|v_*\|} \approx \frac{\|J_*^{-1}\|}{\|v_*\|}\psi(F, v_*, u, \bar{u}) + u. \tag{2.21}$$

Proof. For $i = 0$, the assumptions (2.10) and (2.11) hold with $\nu = \frac{1}{8}$ and $\mu = \frac{1}{8}$ and Theorem 2.2 applies to the first step. Using the values for μ, ν, and the bound (2.18), we find that $G < 1$ so the error contracts if (2.21) does not already hold. Thus, (2.20) is also satisfied with v_0 replaced by $\widehat{v}_1$. The result follows by induction. ∎

Example 1. To illustrate the corollary, we use Newton's method to compute a zero of the polynomial

$$F(v) = (v-1)^{10} - 10^{-8}.$$

At the solution $v_* = 1 - 10^{-0.8} \approx 0.8415$, $|J(v_*)^{-1}| \approx 1.6 \times 10^6$. To increase the rounding errors when computing the residual, we expand $(v-1)^{10}$ as

$$(v-1)^{10} = v^{10} - 10v^9 + 45v^8 - 120v^7 + 210v^6 - 252v^5 + 210v^4 - 120v^3 + 45v^2 - 10v + 1$$

and use this expression to evaluate $F(v)$. For $v \approx 1$ we have $\psi(F, v, u, \bar{u}) \approx 10^3\bar{u}$ (which is roughly the sum of the absolute values of the coefficients in the expansion of $(v-1)^{10}$). Corollary 2.3 predicts that if v_0 is not too far from v_*, the forward error decreases until $|\widehat{v}_{i+1} - v_*|/|v_*| \approx 10^9\bar{u} + u$.

We carried out some numerical experiments in MATLAB, for which the unit roundoff is $u = 2^{-53} \approx 1.1 \times 10^{-16}$. We used the Symbolic Math Toolbox to evaluate $F(v)$ at precision $\bar{u}$. We tried both $\bar{u} = u$ and $\bar{u} = u^{3/2} \approx 3.3 \times 10^{-24}$.[1] The theory predicts

[1] In the BLAST document [2], the term "extended precision" is used for $\bar{u} \leq u^{3/2}$.

limiting accuracy $|\widehat{v}_{i+1} - v_*|/|v_*| \approx 10^{-7}$ if $\bar{u} = u$ and $|\widehat{v}_{i+1} - v_*|/|v_*| \approx 10^{-15}$ if $\bar{u} = u^{3/2}$. For both values of $\bar{u}$, we used two different starting values for v_0, one for which $|v_0 - v_*|/|v_*| > 10^9\bar{u} + u$ and the second one for which the forward error is smaller than the expected limiting accuracy. We plot the behavior of the normwise forward error for $\bar{u} = u$ and $\bar{u} = u^{3/2}$ in Figure 2.1. The results are as predicted by the theory. They also illustrate Wilkinson's remark [27, p. 55]:

> It is perhaps worth remarking that if we start with an approximation to a zero which is appreciably more accurate than the limiting accuracy ... a single iteration will usually spoil this very good approximation and produce one with an error which is typical of the limiting accuracy.

2.3. Residual. We now turn to bounding the residual for a single step of the form (2.6). As before, we write $r = F(v)$ and $J = J(v)$. Note that if $\widehat{v}_* = fl(v_*) = v_* + \Delta v_*$ with $\|\Delta v_*\| \le u\|v_*\|$, then Lemma 2.1 gives

$$F(\widehat{v}_*) = F(v_* + \Delta v_*) = J(v_*)\Delta v_* + \theta, \quad \text{where} \quad \|\theta\| \le \frac{\beta}{2}\|\widehat{v}_* - v_*\|^2.$$

Thus

$$\|F(\widehat{v}_*)\| \le u\|J(v_*)\|\|v_*\| + \frac{\beta}{2}u^2\|v_*\|^2$$

is the best bound we can hope to obtain for the norm of the residual.

THEOREM 2.4. *Assume that there is a v_* such that $F(v_*) = 0$, $J_* = J(v_*)$ is nonsingular, and*

$$\beta\|J_*^{-1}\|\|v - v_*\| \le \mu < 1, \tag{2.22}$$

$$u\|J^{-1}\|\phi(F, v, n, u) \le \nu < 1. \tag{2.23}$$

Let

$$\tau = \beta g\|J_*^{-1}\|,$$

where g is defined in Theorem 2.2. Then

$$\|F(\bar{v})\| \le H\|F(v)\| + h,$$

where

$$H = c_0\left[\mu + \tau + u\kappa(J_*)\right]$$

and

$$h = c_1\left(\mu + \tau + u\kappa(J_*)\right)\psi(F, v, u, \bar{u}) + c_2\left(\mu + \tau + 1\right)u\|J\|\|v\|,$$

with c_0, c_1, and c_2 constants of order 1.

Proof. We have

$$\|J^{-1}E\| \le u\|J^{-1}\|\phi(F, v, n, u) \le \nu < 1 \tag{2.24}$$

using (2.7) and (2.23). Thus, we can apply Theorem 2.2 to deduce that $\bar{v}$ is well defined. Let $\bar{r} = F(\bar{v})$, and define $w \in \mathbb{R}^m$ by $w = \bar{r} - r - J(\bar{v} - v)$. Note that from

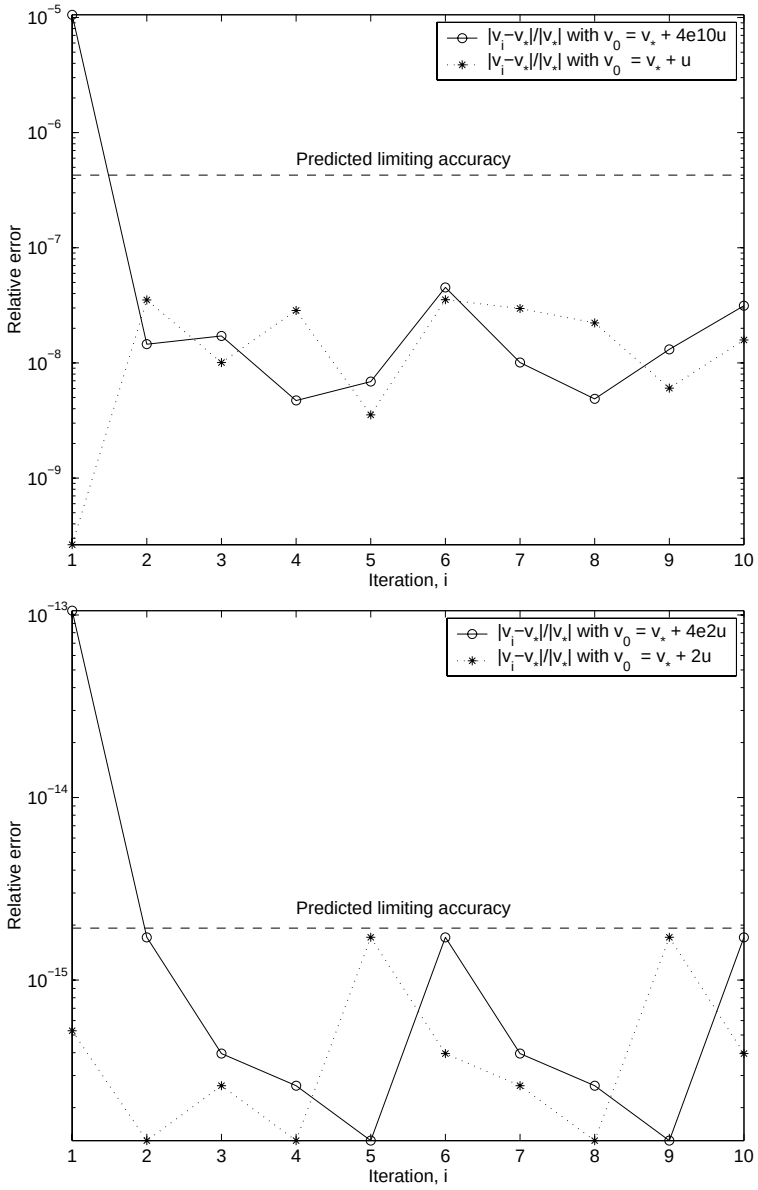

FIG. 2.1. *Behavior of the forward error for $\bar{u} = u$ (top) and $\bar{u} = u^{3/2}$ (bottom).*

(2.6) and (2.8) $\bar{v} - v = -d + \varepsilon$ and $Jd = r + e - Ed$, so that $\bar{r} = r + J(-d + \varepsilon) + w = -e + Ed + J\varepsilon + w$, which yields

$$\begin{aligned} \|\bar{r}\| &\le \|e\| + \|E\|\|d\| + \|J\|\|\varepsilon\| + \|w\| \\ &\le u\|r\| + \psi(F, v, u, \bar{u}) + u\|d\|(\phi(F, v, n, u) + \|J\|) + u\|J\|\|v\| + \|w\|. \end{aligned} \tag{2.25}$$

From (2.12) and (2.13) it follows that

$$\|J\| \le (1 + \mu)\|J_*\|. \tag{2.26}$$

Using (2.17) and (2.24), we have

$$\|d\| \le \|(J+E)^{-1}\|(\|r\|+\|e\|) \le \frac{1}{1-\nu}\|J^{-1}\|\left((1+u)\|r\|+\psi(F,v,u,\bar{u})\right), \tag{2.27}$$

which gives, using (2.14) and (2.26),

$$\begin{aligned} u\|d\|(\phi(F,v,n,u)+\|J\|) \le{}& \frac{1+u}{1-\nu}\left\{u\|J^{-1}\|\phi(F,v,n,u)+\frac{1+\mu}{1-\mu}u\kappa(J_*)\right\}\|r\| \\ &+\frac{1}{1-\nu}\left\{u\|J^{-1}\|\phi(F,v,n,u)+\frac{1+\mu}{1-\mu}u\kappa(J_*)\right\}\psi(F,v,u,\bar{u}). \end{aligned} \tag{2.28}$$

From Lemma 2.1 we have

$$\|w\| \le \frac{\beta}{2}\|\bar{v}-v\|^2. \tag{2.29}$$

First, from (2.6), (2.8), (2.27), and (2.14)

$$\begin{aligned} \|\bar{v}-v\| &\le (1+u)\|d\|+u\|v\| \\ &\le \|J_*^{-1}\|\left(\frac{(1+u)^2}{(1-\mu)(1-\nu)}\|r\|+\frac{1+u}{(1-\mu)(1-\nu)}\psi(F,v,u,\bar{u})\right)+u\|v\|. \end{aligned} \tag{2.30}$$

Second, from the triangle inequality and Theorem 2.2 we have

$$\|\bar{v}-v\| \le (G+1)\|v-v_*\|+g. \tag{2.31}$$

Substituting the product of (2.30) and (2.31) into (2.29) yields

$$\begin{aligned} \|w\| \le{}& \frac{(1+u)^2(G+1)}{2(1-\mu)(1-\nu)}\beta\|J_*^{-1}\|\|v-v_*\|\|r\|+\frac{(1+u)^2}{2(1-\mu)(1-\nu)}\beta\|J_*^{-1}\|g\|r\| \\ &+\frac{(1+u)(G+1)}{2(1-\mu)(1-\nu)}\beta\|J_*^{-1}\|\|v-v_*\|\psi(F,v,u,\bar{u}) \\ &+\frac{(1+u)}{2(1-\mu)(1-\nu)}\beta\|J_*^{-1}\|g\psi(F,v,u,\bar{u}) \\ &+\frac{(G+1)}{2(1-\mu)}\beta\|J_*^{-1}\|\|v-v_*\|u\|J\|\|v\|+\frac{1}{2(1-\mu)}\beta g\|J_*^{-1}\|u\|J\|\|v\|, \end{aligned} \tag{2.32}$$

where the penultimate and last terms on the right-hand side of the inequality are obtained using $\|J\|\|J^{-1}\| \ge 1$ and (2.14). Substituting (2.28) and (2.32) into (2.25) yields

$$\|\bar{r}\| \le H\|r\|+h,$$

with H and h as in the statement of the theorem. ∎

The theorem shows that if the problem is not too ill conditioned, the solver is not too unstable, the approximation of the Jacobian is accurate enough, and v is sufficiently close to the solution, then the norm of the residual reduces after one step of Newton's method in floating point arithmetic. Note that H does not depend on $\bar{u}$ so that, as for the forward error analysis, the use of extended precision for computing the residual has no effect on the rate of convergence of Newton's method. With a careful analysis of the constants in Theorem 2.4 we can derive the following corollary.

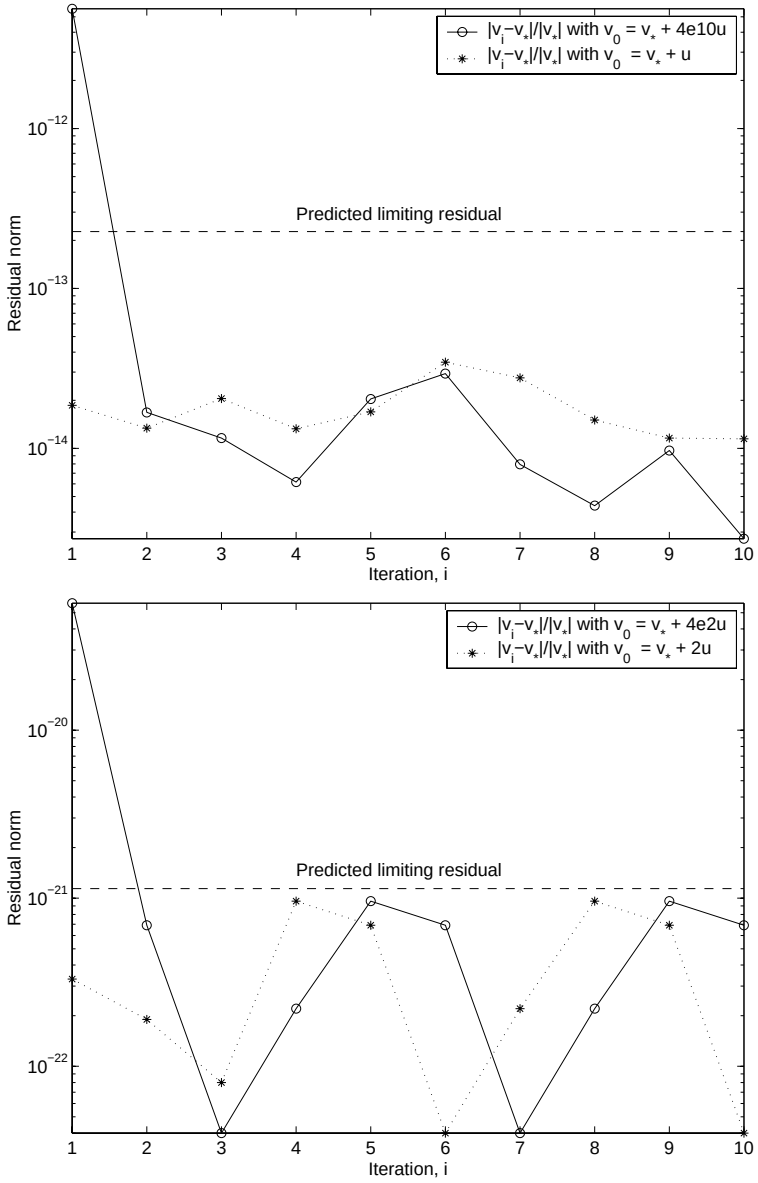

FIG. 2.2. *Behavior of the norm of the residual for $\bar{u} = u$ (top) and for $\bar{u} = u^{3/2}$ (bottom).*

COROLLARY 2.5. *Assume that there is a v_* such that $F(v_*) = 0$, $J_* = J(v_*)$ is nonsingular, and*

$$u\kappa(J_*) < 1/8. \tag{2.33}$$

Assume also that for ϕ in (2.4)

$$u\|J(\widehat{v}_i)^{-1}\|\phi(F, \widehat{v}_i, n, u) < \frac{1}{8} \quad \textit{for all } i \tag{2.34}$$

and that the limiting accuracy $g \approx \|J_*^{-1}\|\psi(F, v_*, u, \bar{u}) + u\|v_*\|$ satisfies $\beta g\|J_*^{-1}\| <$

1/8. *Then, for all* v_0 *such that* $\beta\|J_*^{-1}\|\|v_0 - v_*\| < 1/8$, *the sequence* $\{F(\widehat{v}_i)\}$ *of residual norms generated by Newton's method in floating point arithmetic decreases until*

$$\|F(\widehat{v}_{i+1})\| \approx \psi(F, \widehat{v}_i, u, \bar{u}) + u\|J(\widehat{v}_i)\|\|\widehat{v}_i\|. \tag{2.35}$$

Note that the second term in (2.35) is independent of the accuracy with which the residual is computed.

We consider again Example 1, for which $\psi(F, \widehat{v}_i, u, \bar{u}) \approx 10^3\bar{u}$ and $u\|J(v_*)\|\|v_*\| \approx 10^{-7}u$. As before, we tried both $\bar{u} = u$ and $\bar{u} = u^{3/2} \approx 3.3 \times 10^{-24}$. The theory predicts that

$$\|F(\widehat{v}_i)\| \lesssim \begin{cases} 10^{-13} & \text{if} \quad \bar{u} = u, \\ 10^{-21} & \text{if} \quad \bar{u} = u^{3/2}. \end{cases}$$

We used the same starting values as before. We plot the behavior of $|F(\widehat{v}_i)|$ for $\bar{u} = u$ and $\bar{u} = u^{3/2}$ in Figure 2.2. The results agree well with the predictions.

3. Applications. In this section, we consider several applications. For each of them, we define F and the function ψ and apply our results. We are particularly interested in the effect of mixed precision versus fixed precision for the computation of the residual. The proposed mixed precision BLAS routines (XBLAS) [2] make possible the use of mixed precision in a portable manner.

3.1. Linear systems. We consider the linear system $Ax = b$, where $A \in \mathbb{R}^{n\times n}$ is nonsingular and $b \in \mathbb{R}^n$. Iterative refinement for a computed solution $\widehat{x}$ is simple to describe: compute the residual $r = b - A\widehat{x}$, solve the system $Ad = r$ for the correction d, and form the updated solution $y = \widehat{x} + d$. If necessary, repeat the process with $\widehat{x}$ replaced by y. This process is equivalent to Newton's method with $F(x) = b - Ax$ for which $J(x) = A$ and thus $\beta = 0$.

If the residual $r = F(\widehat{x})$ is computed with the XBLAS routine `GEMV_X` at precision $\bar{u}$, then for ψ in (2.3) we can take

$$\psi(F, \widehat{x}, u, \bar{u}) = \bar{\gamma}_n(\|A\|\|\widehat{x}\| + \|b\|),$$

where $\bar{\gamma}_n$ is defined in (2.5). Corollary 2.3 then yields the following result.

COROLLARY 3.1. *If* $u\kappa(A)$ *is sufficiently less than* 1 *and if the linear system solver is not too unstable, then iterative refinement reduces the relative forward error until*

$$\frac{\|\widehat{x}_i - x\|}{\|x\|} \approx u + \kappa(A)\bar{\gamma}_n.$$

If $\bar{u} = u^2$, *then the relative error is of order* u *provided* $n\kappa(A)u \le 1$.

A backward error of an approximate solution $\widehat{x}$ is a measure of the smallest perturbations ΔA and Δb such that $(A+\Delta A)\widehat{x} = b+\Delta b$. The most popular definition of the normwise backward error is

$$\eta(\widehat{x}) = \min\{\varepsilon : (A + \Delta A)\widehat{x} = b + \Delta b,\ \|\Delta A\| \le \varepsilon\|A\|,\ \|\Delta b\| \le \varepsilon\|b\|\}.$$

It can be shown [21] that

$$\eta(\widehat{x}) = \frac{\|r\|}{\|A\|\|\widehat{x}\| + \|b\|}.$$

Corollary 2.5 thus yields the following result.

COROLLARY 3.2. *Let iterative refinement be applied to the nonsingular linear system $Ax = b$ of order n with $u\kappa(A) < 1/8$ and using a solver satisfying $u\|A^{-1}\|\phi(A, b, n, u) \leq 1/8$. Then the norm of the residual decreases until*

$$\|\widehat{r}_i\| \approx \max(\bar{\gamma}_n, u)(\|A\|\|\widehat{x}\| + \|b\|),$$

so that iterative refinement yields a small normwise backward error $\eta(\widehat{x}) \approx \max(\bar{\gamma}_n, u)$.

Corollaries 3.1 and 3.2 are standard normwise results in the literature [17], [18], [20], [22], [27]. They show that we do not lose anything by using our general analysis.

3.2. Generalized eigenvalue problem. Newton's method and its variants have been considered for improving the accuracy of computed eigenvalues and eigenvectors for the standard eigenvalue problem [10], [7], [8], [23], the singular value problem [9], and refining estimates of invariant subspaces [4], [10]. The error analysis in [10] applies to the standard eigenvalue problem $Ax = \lambda x$ and requires that the problem be scaled ($\|A\| = 1$), that the residual be computed in extended precision, and that the linear solver be stable. A lengthy analysis leads to the conclusion that if the problem is not too ill conditioned and the initial guess is good enough, then their refinement procedure yields a relative error of the order of the working precision.

Here, we consider the generalized eigenvalue problem (GEP)

$$Ax = \lambda Bx \quad \text{with} \quad e_s^T x = 1 \quad \text{for some fixed } s, \tag{3.1}$$

where $A \in \mathbb{R}^{n\times n}, B \in \mathbb{R}^{n\times n}$. Newton-based refinement algorithms for this problem have been proposed [7], [23] but no error analysis has been done.

Define $F : \mathbb{R}^{n+1} \mapsto \mathbb{R}^{n+1}$ by

$$F\left(\begin{bmatrix} x \\ \lambda \end{bmatrix}\right) = \begin{bmatrix} (A - \lambda B)x \\ \alpha e_s^T x - \alpha \end{bmatrix}, \tag{3.2}$$

where $\alpha = \max(\|A\|, \|B\|)$. Then (3.1) can be stated as finding the zeros of $F(v)$, where $v = [x^T, \lambda]^T$. The function F is continuously differentiable in $\mathbb{R}^{n+1}$ with Jacobian

$$J(v) = \begin{bmatrix} A - \lambda B & -Bx \\ \alpha e_s^T & 0 \end{bmatrix}. \tag{3.3}$$

The scalar α is introduced to make F and J scale linearly when A and B are multiplied by a scalar. For all $v, w \in \mathbb{R}^{n+1}$ and any absolute vector norm we have

$$\|J(w) - J(v)\| \leq 2\|B\|\|w - v\|$$

so that J is Lipschitz continuous in $\mathbb{R}^{n+1}$ with constant $\beta = 2\|B\|$.

The next lemma concerns the singularity of J at a zero of F. This result is more general than the one given in [23, p. 120] as it applies to the generalized eigenvalue problem rather than the standard eigenvalue problem and no assumption is made on the nonsingularity of B.

LEMMA 3.3. *Let $v_* = [x_*^T, \lambda_*]^T$ be a zero of F as defined by (3.2) with λ finite. Then $J(v_*)$ is singular if and only if λ_* is a multiple eigenvalue of (A, B).*

Proof. Suppose that $J(v_*)$ is singular. Using the formula (see [13])

$$\det\left(\begin{bmatrix} M & u \\ v_*^T & \mu \end{bmatrix}\right) = \mu \det(M) - v_*^T M^A u,$$

where M^A is the adjugate (or adjoint) of M, we obtain

$$0 = \det(J(v_*)) = \alpha e_s^T (A - \lambda_* B)^A B x_*. \tag{3.4}$$

The adjugate has the property that

$$M^A M = \det(M) I.$$

Define $y^T = e_s^T (A - \lambda_* B)^A$. Then

$$y^T (A - \lambda_* B) = e_s^T \det(A - \lambda_* B) I = 0,$$

because λ_* is an eigenvalue of (A, B). Thus y is a left eigenvector corresponding to λ_*. Using (3.4),

$$y^T B x_* = e_s^T (A - \lambda_* B)^A B x_* = 0.$$

If λ_* were a simple eigenvalue, we would have $y^T B x_* \neq 0$ [1, Thm. 3.2]. So λ_* must be an eigenvalue of multiplicity at least two.

For the converse, suppose that λ_* is a multiple eigenvalue of (A, B). Then, there exists a left eigenvector y corresponding to λ_* that is B-orthogonal to x_*. We have

$$\begin{bmatrix} y^T & 0 \end{bmatrix} \begin{bmatrix} A - \lambda_* B & -B x_* \\ \alpha e_s^T & 0 \end{bmatrix} = 0,$$

which means that $J(v_*)$ is singular. ∎

In exact arithmetic, Theorem 2.2 applies with $E = 0$ and $\nu = u = 0$ so that for all v_0 such that $\|v_0 - v_*\| \leq 1/(4\|B\|\|J_*\|^{-1})$ the Newton iteration is well defined and converges quadratically to zero.

The residual $F(\widehat{v}_i)$ can be computed in mixed precision by the XBLAS routine `GE_SUM_MV`. Then we can take

$$\psi(F, v, u, \bar{u}) = \bar{\gamma}_n(\|A\| + |\lambda|\|B\|)\|x\|. \tag{3.5}$$

COROLLARY 3.4. *Let λ_* be a simple eigenvalue of (A, B), and let x_* be the corresponding eigenvector normalized such that $\|x_*\|_\infty = |x_{*s}| = 1$. Assume that J in* (3.3) *is not too ill conditioned, the linear system solver is not too unstable, and (x_0, λ_0) is a sufficiently good approximation to (x_*, λ_*) so that assumptions* (2.18)–(2.20) *with $\beta = 2\|B\|_\infty$ are satisfied. Then Newton's method for* (3.2) *in floating point arithmetic is well defined and the limiting forward error is bounded by*

$$\frac{\|(\widehat{x}_i^T, \widehat{\lambda}_i) - (x_*^T, \lambda_*)\|_\infty}{\|(x_*^T, \lambda_*)\|_\infty} \lesssim \bar{\gamma}_n \|J(v_*)^{-1}\|_\infty \max(\|A\|_\infty, \|B\|_\infty) + u.$$

If $\bar{u} = u^2$, then

$$\frac{\|(\widehat{x}_i^T, \widehat{\lambda}_i) - (x_*^T, \lambda_*)\|_\infty}{\|(x_*^T, \lambda_*)\|_\infty} \lesssim \gamma_n.$$

Proof. We apply Corollary 2.3 using (3.5) for $\psi(F, v, u, \bar{u})$. We have

$$\begin{aligned} \frac{\|J(v_*)^{-1}\|_\infty}{\|v_*\|_\infty} \psi(F, v_*, u, \bar{u}) &= \frac{\|J(v_*)^{-1}\|_\infty}{\|v_*\|_\infty} \bar{\gamma}_n(\|A\|_\infty + |\lambda_*|\|B\|_\infty)\|x_*\|_\infty \\ &\leq \bar{\gamma}_n \|J(v_*)^{-1}\|_\infty \max(\|A\|_\infty, \|B\|_\infty) \frac{(1 + |\lambda_*|)}{\max(1, |\lambda_*|)} \\ &\leq 2\bar{\gamma}_n \|J(v_*)^{-1}\|_\infty \max(\|A\|_\infty, \|B\|_\infty). \end{aligned}$$

Since $J(v_*)_{n+1,s} = \alpha$, we have $\|J(v_*)\|_\infty \geq \max(\|A\|_\infty, \|B\|_\infty)$. From (2.18), we have $u\kappa(J(v_*)) < 1$ and if $\bar{\gamma}_n \approx nu^2$, then $\bar{\gamma}_n \|J(v_*)^{-1}\|_\infty \lesssim nu \max(\|A\|_\infty, \|B\|_\infty)^{-1}$, which proves the last part of the corollary. ∎

Our result is consistent with the one of Dongarra, Moler, and Wilkinson [10] concerning the standard eigenvalue problem. They showed that their iterative refinement procedure, which is a recasting of Newton's method, yields a forward error of the order of the working precision assuming that $\|A\|_\infty = 1$ and that the residual is computed at precision $\bar{u} = u^2$.

The normwise backward error for an approximate eigenpair $(\widehat{x}, \widehat{\lambda})$ is defined by

$$\eta(\widehat{x}, \widehat{\lambda}) = \min\{\varepsilon : (A + \Delta A)\widehat{x} = \widehat{\lambda}(B + \Delta B)\widehat{x}, \|\Delta A\| \leq \varepsilon\|A\|, \|\Delta B\| \leq \varepsilon\|B\|\},$$

and it can be shown [14], [25] that

$$\eta(\widehat{x}, \widehat{\lambda}) = \frac{\|r\|}{(\|A\| + |\widehat{\lambda}|\|B\|)\|\widehat{x}\|},$$

where $r = A\widehat{x} - \widehat{\lambda}B\widehat{x}$.

COROLLARY 3.5. *Under the same assumptions as in Corollary* 3.4*, Newton's method for* (3.2) *in floating point arithmetic yields a backward error for the* ∞*-norm bounded by*

$$\eta_\infty(\widehat{x}_i, \widehat{\lambda}_i) \lesssim \bar{\gamma}_n + u(3 + |\lambda|) \max\left(\frac{\|A\|_\infty}{\|B\|_\infty}, \frac{\|B\|_\infty}{\|A\|_\infty}\right).$$

Proof. We assume $\|\widehat{x}_i\|_\infty \approx 1$. We have $\psi(F, \widehat{v}_i, u, \bar{u}) \approx \bar{\gamma}_n(\|A\|_\infty + |\widehat{\lambda}_i|\|B\|_\infty)$ and

$$\|\widehat{v}_i\|_\infty \lesssim 1 + |\widehat{\lambda}_i|, \quad \|J(\widehat{v}_i)\|_\infty \lesssim (3 + |\widehat{\lambda}_i|) \max(\|A\|_\infty, \|B\|_\infty),$$

and $(\|A\|_\infty + |\widehat{\lambda}_i|\|B\|_\infty)\|\widehat{x}_i\|_\infty \gtrsim \min(\|A\|_\infty, \|B\|_\infty)(1 + |\widehat{\lambda}_i|)$. Then applying Corollary 2.5 yields the result. ∎

The corollary shows that if $|\lambda| \max(\|A\|_\infty/\|B\|_\infty, \|B\|_\infty/\|A\|_\infty)$ is large, then we cannot guarantee a small backward error. In numerical experiments, we have found that the backward error is small independent of the size of $|\lambda| \max(\|A\|_\infty/\|B\|_\infty, \|B\|_\infty/\|A\|_\infty)$, but we have not been able to prove that this must always be the case.

Note that for the standard eigenvalue problem, $|\lambda_*| \leq 1$ if $\|A\|_\infty = 1$, as was assumed in [10]. Then the eigenpairs refined by Newton's method have a small backward error.

For the GEP, if the problem is scaled and replaced by $\widetilde{A}x = \widetilde{\lambda}Bx$ with $\widetilde{A}$ and $\widetilde{\lambda}$ such that $\|\widetilde{A}\|_\infty = \alpha\|A\|_\infty = \|B\|_\infty$ and $\widetilde{\lambda} = \alpha\lambda$, then, for this problem, the backward error depends only on the size of $|\widetilde{\lambda}|$. A small $|\widetilde{\lambda}|$ ensures a small backward error. If $|\widetilde{\lambda}|$ is large, then we can consider the problem $Bx = \widetilde{\mu}\widetilde{A}x$ for which $|\widetilde{\mu}|$ is small and Corollary 3.5 guarantees that iterative refinement will yield a small backward error.

4. Numerical experiments. We show how iterative refinement can be used to improve the stability of an unstable solver for the symmetric definite generalized eigenvalue problem $Ax = \lambda Bx$, with A symmetric and B symmetric positive definite.

All our tests have been performed with MATLAB for which the working precision is $u = 2^{-53} \approx 1.1 \times 10^{-16}$. We approximate the eigenpairs using the Cholesky-QR method, which consists of the following.

1. Compute the Cholesky factorization $B = GG^T$.
2. Compute $C = G^{-1}AG^{-T}$.
3. Compute the eigendecomposition $W^TCW = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$ using the symmetric QR algorithm.

The matrix $X = G^{-T}W$ is nonsingular and satisfies $X^TBX = I$ and $X^TAX = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$. This algorithm can be unstable. The computed $\widehat{C}$ from step 2 satisfies [3]

$$\widehat{C} = C + \Delta C, \quad \|\Delta C\|_2 \le \gamma_{n^2}\|B^{-1}\|_2\|A\|_2,$$

so if B is ill conditioned, then $\|\Delta C\|_2/\|C\|_2$ can be large, even if the eigenvalue problem itself is well conditioned.

For problem (3.1), the Newton iteration (2.1) can be written as

$$(A - \lambda_i B)\Delta x_{i+1} - \Delta\lambda_{i+1}Bx_i = r_i, \quad e_s^T x_{i+1} = e_s^T x_i = 1, \tag{4.1}$$

where $\Delta x_{i+1} = x_{i+1} - x_i$ and $\Delta\lambda_{i+1} = \lambda_{i+1} - \lambda_i$. As in [10], [23] we note that $e_s^T x_0 = 1$ implies $e_s^T \Delta x_{i+1} = 0$ for $i \ge 0$, and thus the sth column of $A - \lambda_i B$ does not participate in the product with Δx_{i+1}. We can replace the sth column of $A - \lambda_i B$ by $-Bx_i$ and the component s of Δx_{i+1} by $\Delta\lambda_{i+1}$. We define

$$\delta_i = \Delta x_i + \Delta\lambda_i e_s \quad \text{and} \quad M_i = (A - \lambda_i B) - ((A - \lambda_i B)e_s + Bx_i)e_s^T.$$

Then we can rewrite (4.1) as

$$M_i\delta_{i+1} = r_i, \quad \lambda_{i+1} = \lambda_i + e_s^T\delta_{i+1}, \quad x_{i+1} = x_i + \delta_{i+1} - e_s^T\delta_{i+1}e_s. \tag{4.2}$$

Algorithm 4.1 is a straightforward implementation of iteration (4.2).

ALGORITHM 4.1. *Given A, B, and an approximate eigenpair (x, λ) with $\|x\|_\infty = x_s = 1$, this algorithm applies iterative refinement to λ and x:*

repeat until convergence
  $r = \lambda Bx - Ax$ (possibly extended precision used)
  Form M: the matrix $A - \lambda B$ with column s replaced by $-Bx$.
  Factor $PM = LU$ (LU factorization with partial pivoting)
  Solve $M\delta = r$ using the LU factors
  $\lambda = \lambda + \delta_s$; $\delta_s = 0$
  $x = x + \delta$
end

This algorithm is expensive as each iteration requires $O(n^3)$ flops for the factorization of M. If the eigenpairs are approximated by a Cholesky reduction of $A - \lambda B$, then a nonsingular matrix X such that $X^TAX = D = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$ and $X^TBX = I$ is available. Then

$$\begin{aligned} X^T r_i &= X^T M_i \delta_{i+1} \\ &= \left((D - \lambda_i I) - X^T((A - \lambda_i B)e_s + Bx_i)e_s^T X\right) X^{-1}\delta_{i+1}. \end{aligned} \tag{4.3}$$

Defining

$$D_{\lambda_i} = D - \lambda_i I, \quad v_i = X^T((A - \lambda_i B)e_s + Bx_i),$$

$$f = X^T e_s, \quad w_{i+1} = X^{-1}\delta_{i+1}, \quad g_i = X^T r_i,$$

(4.3) becomes

$$(D_{\lambda_i} - v_i f^T) w_{i+1} = g_i. \tag{4.4}$$

The matrix in (4.4) is a rank-one modification of a diagonal matrix. As D_{λ_i} is nearly singular when λ_i approaches the solution λ_*, we cannot use the Sherman–Morrison–Woodbury formula. However, we can define rotations $J_{n-1}, \ldots, J_1$ such that

$$J_1^T \ldots J_{n-1}^T v_i = \pm \|v_i\|_2 e_1,$$

where J_k is a rotation in the $(k, k+1)$ plane. Then $H = J_1^T \ldots J_{n-1}^T D_{\lambda_i}$ is upper Hessenberg, as is the matrix

$$J_1^T \ldots J_{n-1}^T (D_{\lambda_i} - v_i f^T) = H \pm \|v_i\|_2 e_1 f^T = H_1.$$

Using a QR factorization of H_1, the solution of (4.4) can be computed in $O(n^2)$ flops.

ALGORITHM 4.2. *Given A, B, X, and D such that $X^T A X = D$ and $X^T B X = I$ and an approximate eigenpair (x, λ) with $\|x\|_\infty = x_s = 1$, this algorithm applies iterative refinement to λ and x at a cost of $O(n^2)$ flops per iteration.*

repeat until convergence
  $r = \lambda B x - A x$ (possibly extended precision used)
  $D_\lambda = D - \lambda I$
  $d = -Bx - c_{\lambda s}$ where $c_{\lambda s}$ is the sth column of $A - \lambda B$
  $v = X^T d$; $f = X^T e_s$
  Compute Givens rotations J_k in the $(k, k+1)$ plane, such that
    $Q_1^T v := J_1^T \ldots J_{n-1}^T v = \|v\|_2 e_1$
  Compute orthogonal Q_2 such that
    $T = Q_2^T Q_1^T (D_\lambda + v f^T)$ is upper triangular
  $z = Q_2^T Q_1^T X^T r$
  Solve $Tw = z$ for w
  $\delta = Xw$
  $\lambda = \lambda + \delta_s$; $\delta_s = 0$
  $x = x + \delta$
end

When B is ill conditioned, the computed $\widehat{X}$ may be inaccurate, so that $\widehat{X}^T A \widehat{X} = D + \Delta D$, $\widehat{X}^T B \widehat{X} = I + \Delta I$, with possibly large $\|\Delta D\|$ and $\|\Delta I\|$. Then the procedure used in Algorithm 4.2 to solve $M\delta = r$ may be unstable: δ is the exact solution of $(M + \Delta M)\delta = r$ with a possibly large $\|\Delta M\|$. However, the theory shows that allowing some instability in the solver and inaccurate evaluation of the Jacobian (assumptions (2.19) and (2.23)) may affect the rate of convergence of the Newton process but not the limiting accuracy and backward error.

We use the hat notation $(\widehat{x}, \widehat{\lambda})$ for approximate eigenpairs obtained with the Cholesky-QR method and the tilde notation $(\widetilde{x}, \widetilde{\lambda})$ for the refined eigenpairs obtained after a few iterations with Algorithm 4.1 or 4.2 starting with $(\widehat{x}, \widehat{\lambda})$ as initial guess. We need to define several quantities:

$$E_{rel}(\widehat{x}, \widehat{\lambda}) = \|(x, \lambda) - (\widehat{x}, \widehat{\lambda})\|_\infty / \|(x, \lambda)\|_\infty$$

is the relative forward error;

$$\mathrm{cond}(\lambda) = (\|A\|_\infty + |\lambda| \|B\|_\infty) \|x\|_\infty^2 / (|\lambda|\ |y^T B x|)$$

TABLE 4.1
Relative errors, condition numbers, and backward error for Example 1.

	λ_i	$E_{rel}(\widehat{x}_i,\widehat{\lambda}_i)$	$\mathrm{cond}(\lambda_i)$	$\eta(\widehat{x}_i,\widehat{\lambda}_i)$
1	-0.62	6e-5	41	4e-6
2	1.63	6e-5	120	2e-6
3	$9e17$	9e-5	$6e18$	2e-20

TABLE 4.2
Backward error and relative error for the two smallest eigenpairs of Example 1.

			Algorithm 4.1			Algorithm 4.2		
λ_i	η^{est}	E_{rel}^{est}	it	$\eta(\widetilde{x}_i,\widetilde{\lambda}_i)$	$E_{rel}(\widetilde{x}_i,\widetilde{\lambda}_i)$	it	$\eta(\widetilde{x}_i,\widetilde{\lambda}_i)$	$E_{rel}(\widetilde{x}_i,\widetilde{\lambda}_i)$
-0.62	1e-16	1e-14	3	2e-17	2e-16	4	6e-17	4e-16
1.63	1e-16	1e-14	3	3e-17	4e-16	4	4e-17	7e-16

is the condition number of the eigenvalue λ, where y is a left eigenvector corresponding to λ [14];

$$\eta(\widehat{x},\widehat{\lambda}) = \|A\widehat{x} - \widehat{\lambda}B\widehat{x}\|_\infty / \big((\|A\|_\infty + |\widehat{\lambda}|\|B\|_\infty)\|\widehat{x}\|_\infty\big)$$

is the backward error of the approximate eigenpair $(\widehat{x},\widehat{\lambda})$;

$$E_{rel}^{est} = \|J^{-1}\|_\infty \bar{u}(\|A\|_\infty + |\lambda|\|B\|_\infty)\|x\|_\infty / \|(x^T,\lambda)\|_\infty + u$$

is an approximation of the theoretical bound (2.21) for the relative forward error, where the Jacobian matrix J is given by (3.3) and $\psi(F,v,u,\bar{u})$ is given by (3.5) with $\bar{\gamma}_n \approx \bar{u}$; and finally, η^{est} is the theoretical bound of the backward error for the refined eigenpair $(\widetilde{x},\widetilde{\lambda})$ from Corollary 3.5.

Example 1. First we consider

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 4 & 5 \\ 3 & 5 & 6 \end{bmatrix}, \quad G = \begin{bmatrix} .001 & 0 & 0 \\ 1 & .001 & 0 \\ 2 & 1 & 0.001 \end{bmatrix},$$

and $B = GG^T$. This example is used in [12] to illustrate the instability of the Cholesky-QR method when B is ill conditioned. Results are displayed in Table 4.1. The two smallest eigenvalues have a small condition number, but their backward error is large because of the ill conditioning of B ($\kappa_\infty(B) = 7 \times 10^{18}$).

We refined the two smallest eigenvalues using Algorithm 4.1 and Algorithm 4.2 with the approximate eigenpairs as initial guess and the residual computed at working precision ($\bar{u} = u \approx 1.1 \times 10^{-16}$). We terminated the iteration when the norm of the correction stopped decreasing. The results are given in Table 4.2, where it is the number of iterations required for convergence. Algorithm 4.2 uses an unstable solver and therefore requires one more iteration. However, the accuracy and stability are unaffected by this unstable solver. Both algorithms produce refined eigenpairs with a small backward error and a relative error as predicted by the theory.

Example 2. We would like to test the sharpness of the residual bound in Corollary 2.5 and the backward error bound in Corollary 3.5. We consider an example with large $\|J_*\|$, a large ratio $\|A\|_\infty/\|B\|_\infty$, and large eigenvalues. We denote by M the Moler matrix from the Test Matrix Toolbox [15]:

$$m_{ij} = \begin{cases} i \ \text{ if } \ i = j, \\ \min(i,j) - 2 \ \text{ otherwise.} \end{cases}$$

TABLE 4.3
Estimated and computed residuals and backward errors for Example 2.

		Before refinement	From theory		After refinement		
λ_i	cond(λ_i)	$\eta(\widehat{x}_i, \widehat{\lambda}_i)$	$\|r^{est}\|$	η^{est}	$\|r\|$	$\eta(\widetilde{x}_i, \widetilde{\lambda}_i)$	it
7.1e5	2.0	1e-5	3.1e-4	9.1e-5	1.2e-10	5.2e-17	5
5.6e6	9.0	2e-6	1.1e-2	7.3e-4	4.7e-10	4.3e-17	4
2.0e7	29.2	9e-7	1.5e-1	2.6e-3	1.0e-9	2.9e-17	3
3.3e7	48.7	7e-7	4.3e-1	4.3e-3	1.6e-9	2.7e-17	5
4.3e7	62.9	2e-7	7.4e-1	5.6e-3	1.7e-9	2.2e-17	3

TABLE 4.4
Relative error for the computed and refined eigenpairs of Example 3 *using working and double precision in the computation of the residual.*

		Before refinement	After refinement			
			$\bar{u} = u$		$\bar{u} = u^2$	
λ_i	cond(λ_i)	$E_{rel}(\widehat{x}_i, \widehat{\lambda}_i)$	E_{rel}^{est}	$E_{rel}(\widetilde{x}_i, \widetilde{\lambda}_i)$	E_{rel}^{est}	$E_{rel}(\widetilde{x}_i, \widetilde{\lambda}_i)$
2.4e-7	1.8e6	1.3e-8	1.0e-11	2.0e-13	2.2e-16	1.1e-16
2.2e-5	2.0e4	2.1e-8	1.3e-11	7.3e-13	2.2e-16	2.2e-16
8.2e-4	5.3e2	1.0e-9	3.3e-13	1.8e-14	2.2e-16	1.1e-16
1.4e-2	4.0e1	6.9e-11	3.4e-14	2.0e-15	2.2e-16	1.1e-16
2.9e-2	4.6e0	4.3e-11	2.8e-14	5.6e-16	2.2e-16	1.1e-16
1.2e-1	1.5e1	2.6e-11	1.7e-14	5.6e-16	2.2e-16	1.1e-16
1.7e-1	7.4e0	3.6e-11	3.0e-14	1.3e-15	2.2e-16	1.1e-16
3.0e-1	1.1e1	3.0e-11	2.0e-13	2.2e-15	2.2e-16	5.6e-17
3.1e-1	1.2e1	3.4e-11	2.1e-13	7.8e-16	2.2e-16	5.6e-17
9.2e4	3.7e6	1.6e-16	1.5e-9	4.1e-12	2.2e-16	0.0e0

We took $n = 20$, $A = 10^6 I$, and $B = 10^{-2} M$ and computed the approximate eigenpairs using the Cholesky reduction. Instabilities are expected as $\kappa(B) = 2 \times 10^{13}$. All the eigenpairs have a large backward error and a small condition number except the largest one. We refined using Algorithm 4.1. Results for some eigenpairs are given in Table 4.3, where

$$\|r^{est}\| = \bar{u}(\|A\|_\infty + |\lambda|\|B\|_\infty)\|x\|_\infty + u\|J\|_\infty \|(x^T, \lambda)\|_\infty$$

is the theoretical bound (2.35) for the norm of the residual. This example corresponds to the "bad case" where $|\lambda| \max\left(\|A\|/\|B\|, \|B\|/\|A\|\right)$ is large, which explains why the theoretical estimates are so pessimistic. The estimates are sharp when the pair (A, B) is scaled such that $\|A\| = \|B\|$ and the eigenpair is refined on the reverse problem (B, A) if $|\widehat{\lambda}_i|$ is large. We have generated many pairs (A, B) with a large value of $\max\left(\|A\|/\|B\|, \|B\|/\|A\|\right)$ and large eigenvalues, for which the theory predicts a large backward error. For all of them, iterative refinement yields a small backward error as long as the initial guess is good enough for Newton's method to converge.

Example 3. We illustrate how using extended precision in computation of the residual yields a small relative error. Let A be the Prolate matrix of size $n = 10$ of the Test Matrix Toolbox [15], and let B be the Moler matrix. We used the Symbolic Math Toolbox of MATLAB to compute the exact eigenpairs of (A, B) and the Cholesky reduction method to approximate the eigenpairs. We give the results in Table 4.4. We refined using both working precision ($\bar{u} = u$) and double precision ($\bar{u} = u^2$) for the computation of the residual. For eigenpairs such that $E_{rel}(\widehat{x}_i, \widehat{\lambda}_i) > E_{rel}^{est}$, iterative refinement leads to $E_{rel}(\widetilde{x}_i, \widetilde{\lambda}_i) < E_{rel}^{est}$ after two iterations. For the largest eigenvalue, $E_{rel}(\widehat{x}_i, \widehat{\lambda}_i) \ll E_{rel}^{est}$ of $\bar{u} = u$, which means that the approximate eigenpair is appreciably more accurate than the limiting accuracy. In this case, one single step

of iterative refinement is enough to spoil the good initial approximation. If $\bar{u} = u^2$, all the eigenpairs are computed to high relative accuracy as expected from the theory (Corollary 3.4).

For further numerical examples of iterative refinement for the Cholesky-QR method, see [3].

5. Conclusions. We have analyzed Newton's method in floating point arithmetic, allowing for extended precision in computation of the residual, inaccurate evaluation of the Jacobian, and a possibly unstable solver. We estimated the limiting accuracy and the smallest residual norm. We showed that the accuracy with which the residual is computed affects the limiting accuracy. The limiting residual norm depends on two terms, one of them independent of the accuracy used in evaluating the residual.

We applied our results to iterative refinement for the generalized eigenvalue problem. We showed that high accuracy for the refined eigenpairs is guaranteed, under suitable assumptions, if twice the working precision is used for the computation of the residual. We also showed that if the pair (A, B) is well balanced ($\|A\| \approx \|B\|$), working precision in evaluating the residual is enough for iterative refinement to yield a small backward error.

Finally, we examined in detail how iterative refinement can be used to improve the forward and backward error of computed eigenpairs for the symmetric definite GEP. We used two refinement algorithms, one of them with an unstable solver. We confirmed that the unstable solver affects the convergence but not the limiting accuracy and backward error. In practice, the assumption that the pair (A, B) is well balanced does not seem to be necessary. We have not been able to generate an example for which iterative refinement fails to yield a small backward error for pairs (A, B) for which $\max(\|A\|/\|B\|, \|B\|/\|A\|)$ is large. This suggests that the bound of Corollary 3.5 is pessimistic. Deriving a sharper bound remains an open problem.

In future work, we plan to investigate iterative refinement for the quadratic eigenvalue problem, for which there are no proven backward stable algorithms [25].

Acknowledgments. I thank the referees for valuable suggestions that improved the paper.

REFERENCES

[1] A. L. Andrew, K.-W. E. Chu, and P. Lancaster, *Derivatives of eigenvalues and eigenvectors of matrix functions*, SIAM J. Matrix Anal. Appl., 14 (1993), pp. 903–926.

[2] *BLAS Technical Forum Standard*, International Journal of High Performance Computing Applications, to appear. Available online at http://www.netlib.org/blas/blast-forum/.

[3] P. I. Davies, N. J. Higham, and F. Tisseur, *Analysis of the Cholesky Method with Iterative Refinement for Solving the Symmetric Definite Generalized Eigenproblem*, Numerical Analysis Report No. 360, Manchester Centre for Computational Mathematics, Manchester, UK, 2000.

[4] J. W. Demmel, *Three methods for refining estimates of invariant subspaces*, Computing, 38 (1987), pp. 43–57.

[5] J. E. Dennis, Jr. and R. B. Schnabel, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, Prentice-Hall, Englewood Cliffs, NJ, 1983.

[6] J. E. Dennis, Jr. and H. F. Walker, *Inaccuracy in quasi-Newton methods: Local improvement theorems*, Math. Programming Stud., 22 (1984), pp. 70–85.

[7] J. J. Dongarra, *Improving the accuracy of computed matrix eigenvalues*, Preprint ANL-80-84, Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, IL, 1980.

[8] J. J. Dongarra, *Algorithm* 589 *SICEDR: A FORTRAN subroutine for improving the accuracy of computed matrix eigenvalues*, ACM Trans. Math. Software, 8 (1982), pp. 371–375.
[9] J. J. Dongarra, *Improving the accuracy of computed singular values*, SIAM J. Sci. Statist. Comput., 4 (1983), pp. 712–719.
[10] J. J. Dongarra, C. B. Moler, and J. H. Wilkinson, *Improving the accuracy of computed eigenvalues and eigenvectors*, SIAM J. Numer. Anal., 20 (1983), pp. 23–45.
[11] A. R. Ghavimi and A. J. Laub, *Backward error, sensitivity, and refinement of computed solutions of algebraic Riccati equations*, Numer. Linear Algebra Appl., 2 (1995), pp. 29–49.
[12] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 3rd ed., Johns Hopkins University Press, Baltimore, MD, 1996.
[13] H. V. Henderson and S. R. Searle, *On deriving the inverse of a sum of matrices*, SIAM Rev., 23 (1981), pp. 53–60.
[14] D. J. Higham and N. J. Higham, *Structured backward error and condition of generalized eigenvalue problems*, SIAM J. Matrix Anal. Appl., 20 (1998), pp. 493–512.
[15] N. J. Higham, *The Test Matrix Toolbox for* Matlab *(version* 3.0*)*, Numerical Analysis Report No. 276, Manchester Centre for Computational Mathematics, Manchester, UK, 1995.
[16] N. J. Higham, *Accuracy and Stability of Numerical Algorithms*, SIAM, Philadelphia, 1996.
[17] N. J. Higham, *Iterative refinement for linear systems and LAPACK*, IMA J. Numer. Anal., 17 (1997), pp. 495–509.
[18] M. Jankowski and H. Woźniakowski, *Iterative refinement implies numerical stability*, BIT, 17 (1977), pp. 303–311.
[19] P. Lancaster, *Error analysis for the Newton-Raphson method*, Numer. Math., 9 (1966), pp. 55–68.
[20] C. B. Moler, *Iterative refinement in floating point*, J. Assoc. Comput. Mach., 14 (1967), pp. 316–321.
[21] J. L. Rigal and J. Gaches, *On the compatibility of a given solution with the data of a linear system*, J. Assoc. Comput. Mach., 14 (1967), pp. 543–548.
[22] R. D. Skeel, *Iterative refinement implies numerical stability for Gaussian elimination*, Math. Comp., 35 (1980), pp. 817–832.
[23] H. J. Symm and J. H. Wilkinson, *Realistic error bounds for a simple eigenvalue and its associated eigenvector*, Numer. Math., 35 (1980), pp. 113–126.
[24] R. A. Tapia, *The Kantorovich theorem for Newton's method*, Amer. Math. Monthly, 78 (1971), pp. 389–392.
[25] F. Tisseur, *Backward error and condition of polynomial eigenvalue problems*, Linear Algebra Appl., 309 (2000), pp. 339–361.
[26] S. Wang and S. Zhao, *An algorithm for* $Ax = \lambda Bx$ *with symmetric and positive-definite* A *and* B, SIAM J. Matrix Anal. Appl., 12 (1991), pp. 654–660.
[27] J. H. Wilkinson, *Rounding Errors in Algebraic Processes*, Notes on Applied Science No. 32, Her Majesty's Stationery Office, London, 1963. Also published by Prentice-Hall, Englewood Cliffs, NJ, 1963. Reprinted by Dover, New York, 1994.
[28] H. Woźniakowski, *Numerical stability for solving nonlinear equations*, Numer. Math., 27 (1977), pp. 373–390.
[29] T. J. Ypma, *The effect of rounding errors on Newton-like methods*, IMA J. Numer. Anal., 3 (1983), pp. 109–118.
[30] T. J. Ypma, *Local convergence of inexact Newton methods*, SIAM J. Numer. Anal., 21 (1984), pp. 583–590.