ISSN 1360-1725

THE UNIVERSITY of MANCHESTER

UMIST

Iterative Refinement and LAPACK

N. J. Higham

Numerical Analysis Report No. 277

September 1995

Manchester Centre for Computational Mathematics
Numerical Analysis Reports

DEPARTMENTS OF MATHEMATICS

Reports available from:
Department of Mathematics
University of Manchester
Manchester M13 9PL
England

And over the World-Wide Web from URLs
http://www.ma.man.ac.uk/MCCM/MCCM.html
ftp://vtx.ma.man.ac.uk/pub/narep

Iterative Refinement and LAPACK

Nicholas J. Higham*

September 29, 1995

Abstract

The technique of iterative refinement for improving the computed solution to a linear system was used on desk calculators and computers in the 1940s and has remained popular. In the 1990s iterative refinement is well supported in software libraries, notably in LAPACK. Although the behaviour of iterative refinement in floating point arithmetic is reasonably well understood, the existing theory is not sufficient to justify the use of fixed precision iterative refinement in all the LAPACK routines in which it is implemented. We present analysis that provides the theoretical support needed for LAPACK. The analysis covers both mixed and fixed precision iterative refinement with an arbitrary number of iterations, makes only a general assumption on the underlying solver, and is relatively short. We identify some remaining open problems.

Key words. iterative refinement, rounding error analysis, backward error, condition number, LAPACK

AMS subject classifications. primary 65F05, 65G05

1 Introduction

The technique of iterative refinement for improving the computed solution to a linear system was probably first used in a computer program by Wilkinson in 1948, during the design and building of the ACE computer at the National Physical Laboratory [15]. Iterative refinement has achieved wide use ever since, and is exploited, for example, by most of the linear system expert drivers in LAPACK [1].

The refinement process for a computed solution $\widehat{x}$ to $Ax = b$, where $A \in \mathbb{R}^{n\times n}$ is nonsingular, 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 there is not

*Department of Mathematics, University of Manchester, Manchester, M13 9PL, England (na.nhigham@na-net.ornl.gov). This work was supported by Engineering and Physical Sciences Research Council grants GR/H5213 and GR/H/94528.

a sufficient improvement in passing from $\widehat{x}$ to y the process can be repeated, with $\widehat{x}$ replaced by y.

Intuition suggests that, since the residual r contains the crucial information that enables $\widehat{x}$ to be improved, r should be computed as accurately as possible. In the early application and analysis of iterative refinement r was computed in extended precision and then rounded to working precision. This *mixed precision* iterative refinement was analyzed by Wilkinson [16] and Moler [10]; they showed that, provided A is not too ill conditioned, it produces a computed solution correct to working precision. Mixed precision iterative refinement contrasts with *fixed precision* iterative refinement, in which r is formed entirely in the working precision. In the late 1970s Skeel [12] proved that, under certain conditions, just one step of fixed precision iterative refinement is sufficient to yield a small componentwise relative backward error for Gaussian elimination with partial pivoting (GEPP) (the componentwise relative backward error is defined below); Jankowski and Woźniakowski [9] had earlier shown that, again with certain provisos, an *arbitrary* linear equation solver is made normwise backward stable by fixed precision iterative refinement (possibly with more than one iteration).

Skeel's analysis of fixed precision iterative refinement was generalized by Higham [4] to an arbitrary linear equation solver satisfying certain stability assumptions. This general analysis can be used to show that "one step is enough" for GEPP and for solvers based on QR factorization computed by any of the standard methods; the analysis also has applications to methods for solving the least squares problem. Unfortunately, Higham's analysis does not yield any useful conclusions about the componentwise relative backward error resulting from fixed precision iterative refinement applied with the Cholesky factorization or the diagonal pivoting method. In LAPACK, both these methods are implemented with the option of performing fixed precision iterative refinement, but there is no existing theory to prove that a small componentwise relative backward error will usually be achieved.

The purpose of this work is to present a general analysis that fills the gaps in our understanding of iterative refinement and yields positive conclusions for the Cholesky factorization and the diagonal pivoting method.

In the rest of the introduction we present the required notions of stability and conditioning.

We recall the definition of *componentwise backward error* for an approximate solution y to a linear system $Ax = b$:

$$\omega_{E,f}(y) = \min\{\epsilon : (A + \Delta A)y = b + \Delta b, \ \ |\Delta A| \le \epsilon E, \ \ |\Delta b| \le \epsilon f\}, \tag{1.1}$$

where E and f are nonnegative matrices of tolerances. For $E = |A|$ and $f = |b|$ we obtain the *componentwise relative backward error*. A computationally simple formula exists for $\omega_{E,f}(y)$, as shown in the following result. We adopt the convention that $\xi/0$ is interpreted as zero if $\xi = 0$ and infinity otherwise.

Theorem 1.1 *The componentwise backward error is given by*

$$\omega_{E,f}(y) = \max_i \frac{|r_i|}{(E|y| + f)_i}, \tag{1.2}$$

where $r = b - Ay$.

Proof. See Oettli and Prager [11], or Higham [7, Th. 7.3]. $\square$

We introduce the condition numbers

$$\begin{aligned} \operatorname{cond}(A, x) &:= \frac{\| \, |A^{-1}||A||x| \, \|_\infty}{\|x\|_\infty}, \\ \operatorname{cond}(A) &:= \operatorname{cond}(A, e) = \| \, |A^{-1}||A| \, \|_\infty \le \|A^{-1}\|_\infty \|A\|_\infty = \kappa_\infty(A), \end{aligned}$$

where $e = [1, 1, \ldots, 1]^T$. The term "condition number" is used advisedly here. If we define the componentwise condition number

$$\begin{aligned} \operatorname{cond}_{E,f}(A, x) := \lim_{\epsilon \to 0} \sup \Big\{ \frac{\|\Delta x\|_\infty}{\epsilon \|x\|_\infty} : (A + \Delta A)(x + \Delta x) = b + \Delta b, \\ |\Delta A| \le \epsilon E, \quad |\Delta b| \le \epsilon f \Big\}, \end{aligned}$$

then $\operatorname{cond}(A, x)$ is within a factor 2 of $\operatorname{cond}_{|A|,|b|}(A, x)$, and $\operatorname{cond}(A)$ differs from the condition number corresponding to $E = |A|ee^T$ and $f = |b|$ by at most a factor $2n$ [7, Problem 7.6].

We will need a corollary of Theorem 1.1 in which x replaces y in the expression $E|y| + f$. First, we state a trivial lemma, which involves a function ψ that measures how badly a vector is scaled.

Lemma 1.2 *For* $B \in \mathbb{R}^{n \times n}$ *and* $y \in \mathbb{R}^n$ *we have*

$$|B||y| \le \|B\|_\infty \psi(y) |y|,$$

where

$$\psi(y) = \frac{\max_i |y_i|}{\min_i |y_i|}.$$

Proof. We have

$$|B||y| \le \|B\|_\infty \|y\|_\infty e \le \|B\|_\infty \psi(y) |y|. \qquad \square$$

Corollary 1.3 *The componentwise backward error satisfies*

$$\frac{\theta}{1 + \| \, E|A^{-1}| \, \|_\infty \psi(E|x| + f)\theta} \le \omega_{E,f}(y) \le \frac{\theta}{1 - \| \, E|A^{-1}| \, \|_\infty \psi(E|x| + f)\theta}, \tag{1.3}$$

where

$$\theta = \max_i \frac{|r_i|}{(E|x| + f)_i},$$

$r = b - Ay$, *and the denominators are assumed to be positive.*

Proof. Dividing numerator and denominator in (1.2) by $(E|x| + f)_i$, and using the inequality $E|y| \geq E|x| - E|x - y| \geq E|x| - E|A^{-1}||r|$, we obtain the bound

$$\omega_{E,f}(y) \leq \left(\max_i \frac{|r_i|}{(E|x| + f)_i} \right) \Big/ \left(1 - \max_i \frac{(E|A^{-1}||r|)_i}{(E|x| + f)_i} \right).$$

Using Lemma 1.2 we have

$$\begin{aligned} E|A^{-1}||r| &\leq \theta\, E|A^{-1}|(E|x| + f) \\ &\leq \theta \, \| \, E|A^{-1}| \, \|_\infty \psi(E|x| + f)(E|x| + f), \end{aligned}$$

which gives the upper bound. The lower bound is proved similarly. □

Finally, we note that our backward error results in §4 are cast in terms of $\psi(|b| + |A||x|)$. This is related to the quantity $\psi(|A||x|)$ that appears in the analyses of Skeel [12] and Higham [4] by the inequalities

$$\frac{1}{2}\psi(|A||x|) \leq \psi(|b| + |A||x|) \leq 2\psi(|A||x|).$$

2 Basics

We work with the standard model of floating point arithmetic:

$$fl(x \text{ op } y) = (x \text{ op } y)(1 + \delta), \qquad |\delta| \leq u, \quad \text{op} = +, -, *, /, \tag{2.1}$$

where u is the unit roundoff. (In fact, our results hold, with minor changes to the constants, under a weaker model that accommodates machines without a guard digit [7, §2.4].) We will make use of the constant

$$\gamma_n = \frac{nu}{1 - nu}.$$

As this notation suggests, we assume that $nu < 1$.

Consider a linear system $Ax = b$, where $A \in \mathbb{R}^{n \times n}$ is nonsingular. We suppose that our solver produces a computed solution satisfying

$$(A + \Delta A)\widehat{x} = b, \qquad |\Delta A| \leq uW, \tag{2.2}$$

where W is a nonnegative matrix depending on A, n and u (but not on b). We invariably have $W \geq |A|$. Note that at this stage we make no assumption about the size or structure of W. All standard direct solvers satisfy (2.2), and iterative solvers may, too, depending on the convergence test (see, for, example, [8]). Although backward error results for solution of $Ax = b$ by QR factorization methods are usually stated with a perturbation of b, these results can be reworked so that only A is perturbed (see [7, §18.3]). The advantage of perturbing only A in (2.2) is that we obtain an algebraically simpler analysis of iterative refinement.

Inevitably, our analysis requires A not to be too ill conditioned. We make an initial assumption that

$$u\| \, |A^{-1}|W \, \|_\infty < \frac{1}{2}, \tag{2.3}$$

which guarantees that $A + \Delta A$ in (2.2) is nonsingular and enables us to bound $(I - u|A^{-1}|W)^{-1}$ in §4.

To make the analysis as general as possible we allow for the use of extended precision in calculating the residual. We define $x_1 = \widehat{x}$ (equivalently, $x_0 = 0$) and consider the following iterative refinement process: $r_i = b - Ax_i$ (precision $\overline{u}$), solve $Ad_i = r_i$ (precision u), $x_{i+1} = x_i + d_i$ (precision u), $i = 1, 2, \ldots$. For traditional iterative refinement, $\overline{u} = u^2$.

There are two stages in the calculation of r_i. First, $s_i = fl(b - A\widehat{x}_i) = b - A\widehat{x}_i + \Delta s_i$ is formed in the (possibly) extended precision $\overline{u}$. Standard results [7, §3.5] show that $|\Delta s_i| \le \overline{\gamma}_{n+1}(|b| + |A||\widehat{x}_i|)$, where $\overline{\gamma}_k \le k\overline{u}/(1 - k\overline{u})$. Second, the residual is rounded to the working precision: $\widehat{r}_i = fl(s_i) = s_i + f_i$, where $|f_i| \le u|s_i|$. Hence

$$\widehat{r}_i = r_i + \Delta r_i, \qquad |\Delta r_i| \le u|r_i| + (1 + u)\overline{\gamma}_{n+1}(|b| + |A||\widehat{x}_i|). \tag{2.4}$$

3 Forward Error Analysis

We begin by analyzing the behaviour of the forward error of $\widehat{x}_i$, namely, $\|x - \widehat{x}_i\|_\infty / \|x\|_\infty$. The analysis in this section is a slightly rewritten version of the analysis in Higham [7, §11.1].

By writing $\widehat{x}_i = x + (\widehat{x}_i - x)$ and $r_i = A(x - \widehat{x}_i)$ we obtain from (2.4) the bound

$$|\Delta r_i| \le [u + (1 + u)\overline{\gamma}_{n+1}]|A||x - \widehat{x}_i| + 2(1 + u)\overline{\gamma}_{n+1}|A||x|. \tag{3.1}$$

For the computation of d_i we have, by (2.2), $(A + \Delta A_i)\widehat{d}_i = \widehat{r}_i$, where $|\Delta A_i| \le uW$. Now write

$$(A + \Delta A_i)^{-1} = \bigl(A(I + A^{-1}\Delta A_i)\bigr)^{-1} =: (I + F_i)A^{-1},$$

where

$$|F_i| \le u|A^{-1}|W + O(u^2). \tag{3.2}$$

Hence

$$\widehat{d}_i = (I + F_i)A^{-1}\widehat{r}_i = (I + F_i)(x - \widehat{x}_i + A^{-1}\Delta r_i). \tag{3.3}$$

The updated vector satisfies

$$\begin{aligned}\widehat{x}_{i+1} &= \widehat{x}_i + \widehat{d}_i + \Delta\widehat{x}_i,\\ |\Delta\widehat{x}_i| &\le u|\widehat{x}_i + \widehat{d}_i| \le u(|x-\widehat{x}_i| + |x| + |\widehat{d}_i|).\end{aligned}$$

Using (3.3) we have

$$\widehat{x}_{i+1} - x = F_i(x-\widehat{x}_i) + (I+F_i)A^{-1}\Delta r_i + \Delta\widehat{x}_i.$$

Hence

$$\begin{aligned}|\widehat{x}_{i+1} - x| &\le |F_i||x-\widehat{x}_i| + (I+|F_i|)|A^{-1}||\Delta r_i| + u|x-\widehat{x}_i| + u|x| + u|\widehat{d}_i|\\ &\le |F_i||x-\widehat{x}_i| + (I+|F_i|)|A^{-1}||\Delta r_i| + u|x-\widehat{x}_i| + u|x|\\ &\qquad + u(I+|F_i|)(|x-\widehat{x}_i| + |A^{-1}||\Delta r_i|)\\ &= \big((1+u)|F_i| + 2uI\big)|x-\widehat{x}_i| + (1+u)(I+|F_i|)|A^{-1}||\Delta r_i| + u|x|.\end{aligned}$$

Substituting the bound for $|\Delta r_i|$ from (3.1) gives

$$\begin{aligned}|\widehat{x}_{i+1} - x| \le\ & \big((1+u)|F_i| + 2uI\big)|x-\widehat{x}_i|\\ &\qquad + (1+u)\big(u + (1+u)\overline{\gamma}_{n+1}\big)(I+|F_i|)|A^{-1}||A||x-\widehat{x}_i|\\ &\qquad + 2(1+u)^2\overline{\gamma}_{n+1}(I+|F_i|)|A^{-1}||A||x| + u|x|\\ =:\ & G_i|x-\widehat{x}_i| + g_i. \end{aligned} \tag{3.4}$$

Using (3.2), we estimate

$$\begin{aligned}G_i &\approx |F_i| + (u+\overline{\gamma}_{n+1})(I+|F_i|)|A^{-1}||A|\\ &\lesssim u|A^{-1}|W + (u+\overline{\gamma}_{n+1})(I+u|A^{-1}|W)|A^{-1}||A|,\\ g_i &\approx 2\overline{\gamma}_{n+1}(I+|F_i|)|A^{-1}||A||x| + u|x|\\ &\lesssim 2\overline{\gamma}_{n+1}(I+u|A^{-1}|W)|A^{-1}||A||x| + u|x|.\end{aligned}$$

Recall that we are assuming (2.3) holds. As long as A is not too ill conditioned ($\mathrm{cond}(A)$ is not too large) we have $\|G_i\|_\infty < 1$, which means that the error contracts until we reach a point at which the g_i term becomes significant. The limiting normwise accuracy, that is, the minimum size of $\|x-\widehat{x}_i\|_\infty/\|x\|_\infty$, is roughly $\|g_i\|_\infty/\|x\|_\infty \approx 2n\overline{u}\,\mathrm{cond}(A,x) + u$. Moreover, if $2n\overline{u}(I+u|A^{-1}|W)|A^{-1}||A||x| \le \mu u|x|$ for some μ, then we can expect to obtain a componentwise relative error of order μu, that is, $\min_i |x-\widehat{x}_i| \lesssim \mu u|x|$. Note that G_i is essentially independent of $\overline{u}$, which suggests that the rate of convergence of mixed and fixed precision iterative refinement will be similar; it is only the limiting accuracy that differs.

In the traditional use of iterative refinement, $\overline{u} = u^2$, and one way to summarize our findings is as follows.

Theorem 3.1 (Mixed precision iterative refinement) *Let iterative refinement be applied to the nonsingular linear system $Ax = b$, using a solver satisfying* (2.2) *and with residuals computed in double the working precision. Let $\eta = u\| \, |A^{-1}|(|A| + W) \, \|_\infty$. Then, provided η is sufficiently less than* 1*, iterative refinement reduces the forward error by a factor approximately η at each stage, until $\|x - \widehat{x}_i\|_\infty / \|x\|_\infty \approx u$.* □

For LU factorization we can take

$$uW \equiv 2\gamma_n |\widehat{L}||\widehat{U}| \tag{3.5}$$

[7, Th. 9.3], where $\widehat{L}$ and $\widehat{U}$ are the computed LU factors. In this case Theorem 3.1 is stronger than the standard results in the literature, which have $\kappa_\infty(A)u$ in place of $\eta \approx u\| \, |A^{-1}|(|A|+2n|\widehat{L}||\widehat{U}|) \, \|_\infty$. For we can have $\eta \ll \kappa_\infty(A)u$, since η is independent of the row scaling of A (modulo changes in the pivot sequence). For example, if $|\widehat{L}||\widehat{U}| \approx |A|$ then $\eta \approx 2n \operatorname{cond}(A)u$, and $\operatorname{cond}(A)$ can be arbitrarily smaller than $\kappa_\infty(A)$.

Consider now fixed precision iterative refinement, in which $\overline{u} = u$. We have the following analogue of Theorem 3.1, which refutes claims in some textbooks that for iterative refinement to improve the accuracy it is necessary to compute the residual in extra precision.

Theorem 3.2 (Fixed precision iterative refinement) *Let iterative refinement in fixed precision be applied to the nonsingular linear system $Ax = b$ of order n, using a solver satisfying* (2.2)*. Let $\eta = u\| \, |A^{-1}|(|A| + W) \, \|_\infty$. Then, provided η is sufficiently less than* 1*, iterative refinement reduces the forward error by a factor approximately η at each stage, until $\|x - \widehat{x}_i\|_\infty / \|x\|_\infty \lesssim 2n \operatorname{cond}(A, x)u$.* □

The key difference between mixed and fixed precision iterative refinement is that in the latter case a relative error of order u is no longer ensured. But we do have a relative error bound of order $\operatorname{cond}(A, x)u$. This is a stronger bound than holds for the original computed solution $\widehat{x}$, for which we can say only that

$$\frac{\|x - \widehat{x}\|_\infty}{\|x\|_\infty} \lesssim u \frac{\| \, |A^{-1}|W|x| \, \|_\infty}{\|x\|_\infty}.$$

In fact, a relative error bound of order $\operatorname{cond}(A, x)u$ is the best we can possibly expect if we do not use higher precision, because it corresponds to the uncertainty introduced by making componentwise relative perturbations to A of size u. This level of uncertainty is often present in the problem as it is given, because of errors in computing A or in rounding its elements to floating point form.

4 Backward Error Analysis

We now turn to the backward error. The analysis in this section generalizes that of Skeel [12] by applying to any solver satisfying (2.2), rather than just GEPP, and it generalizes the analysis of Higham [4] by applying to both mixed and fixed precision iterative refinement with an arbitrary number of steps, rather than just one step of fixed precision refinement.

In the analysis we endeavour to obtain bounds containing terms that are multiples of $|A||x|$. To this end, we make frequent use of the following trivial inequality:

$$\begin{aligned} |A||\widehat{x}_i| &\le |A||x| + |A||x-\widehat{x}_i| \\ &\le |A||x| + |A||A^{-1}||b-A\widehat{x}_i| \\ &= |A||x| + |A||A^{-1}||r_i|. \end{aligned} \tag{4.1}$$

For later use we note that, from (2.4),

$$\begin{aligned} |\Delta r_i| &\le u|r_i| + (1+u)\overline{\gamma}_{n+1}(|b| + |A||\widehat{x}_i|) \\ &\le \big(uI + (1+u)\overline{\gamma}_{n+1}|A||A^{-1}|\big)|r_i| + (1+u)\overline{\gamma}_{n+1}(|b| + |A||x|) \\ &= C_1|r_i| + (1+u)\overline{\gamma}_{n+1}(|b| + |A||x|), \end{aligned} \tag{4.2}$$

where

$$C_1 = uI + (1+u)\overline{\gamma}_{n+1}|A||A^{-1}|.$$

Then

$$|\widehat{r}_i| \le M_1|r_i| + (1+u)\overline{\gamma}_{n+1}(|b| + |A||x|), \tag{4.3}$$

where

$$M_1 = I + C_1.$$

For the solution of the correction equation we have

$$A\widehat{d}_i = \widehat{r}_i + f_1, \qquad |f_1| \le uW|\widehat{d}_i|, \tag{4.4}$$

and the updated vector satisfies

$$\widehat{x}_{i+1} = \widehat{x}_i + \widehat{d}_i + f_2, \qquad |f_2| \le u|\widehat{x}_i + \widehat{d}_i|. \tag{4.5}$$

We have

$$\begin{aligned} b - A\widehat{x}_{i+1} &= b - A\widehat{x}_i - A\widehat{d}_i - Af_2 \\ &= \widehat{r}_i - \Delta r_i - A\widehat{d}_i - Af_2 \\ &= -f_1 - \Delta r_i - Af_2. \end{aligned}$$

Hence, using (4.4), (4.2), (4.5), and (4.1) we have

$$
\begin{aligned}
|b - A\widehat{x}_{i+1}| &\le uW|\widehat{d}_i| + C_1|r_i| + (1+u)\overline{\gamma}_{n+1}(|b| + |A||x|) + u|A|(|\widehat{x}_i| + |\widehat{d}_i|) \\
&\le uW|\widehat{d}_i| + (C_1 + u|A||A^{-1}|)|r_i| \\
&\qquad + (1+u)\overline{\gamma}_{n+1}(|b| + |A||x|) + u|A|(|x| + |\widehat{d}_i|) \\
&= u(W + |A|)|\widehat{d}_i| + C_2|r_i| + (1+u)\overline{\gamma}_{n+1}(|b| + |A||x|) + u|A||x|, \quad (4.6)
\end{aligned}
$$

where

$$C_2 = C_1 + u|A||A^{-1}|.$$

Our aim is to bound $\omega_{|A|,|b|}(\widehat{x}_{i+1})$ using Corollary 1.3, so we need to bound $(W + |A|)|\widehat{d}_i|$ by multiples of $|r_i|$ and $|b| + |A||x|$. From (4.4) and (4.3),

$$
\begin{aligned}
|\widehat{d}_i| &\le |A^{-1}|(|\widehat{r}_i| + uW|\widehat{d}_i|) \\
&\le |A^{-1}|\big(M_1|r_i| + (1+u)\overline{\gamma}_{n+1}(|b| + |A||x|) + uW|\widehat{d}_i|\big), \qquad (4.7)
\end{aligned}
$$

that is,

$$(I - u|A^{-1}|W)|\widehat{d}_i| \le |A^{-1}|\big(M_1|r_i| + (1+u)\overline{\gamma}_{n+1}(|b| + |A||x|)\big).$$

In view of the assumption (2.3) we have

$$|\widehat{d}_i| \le M_2|A^{-1}|\big(M_1|r_i| + (1+u)\overline{\gamma}_{n+1}(|b| + |A||x|)\big), \qquad (4.8)$$

where

$$M_2 = (I - u|A^{-1}|W)^{-1} \ge 0, \qquad \|M_2\|_\infty \le 2.$$

Substituting into (4.6) gives

$$
\begin{aligned}
|r_{i+1}| &\le u(W + |A|)M_2|A^{-1}|\big(M_1|r_i| + (1+u)\overline{\gamma}_{n+1}(|b| + |A||x|)\big) + C_2|r_i| \\
&\qquad + (1+u)\overline{\gamma}_{n+1}(|b| + |A||x|) + u|A||x| \\
&\le \big(C_2 + u(W + |A|)M_2|A^{-1}|M_1\big)|r_i| \\
&\qquad + \big[(u + (1+u)\overline{\gamma}_{n+1})I + u(1+u)\overline{\gamma}_{n+1}(W + |A|)M_2|A^{-1}|\big](|b| + |A||x|) \\
&=: G|r_i| + g. \qquad (4.9)
\end{aligned}
$$

Note that $\|C_2\|_\infty = O(u\,\mathrm{cond}(A^{-1}))$ and $M_i = I + O(u)$, $i = 1{:}2$. As long as A is not too ill conditioned and the solver is not too unstable we have $\|G\|_\infty < 1$. Then, solving the recurrence, we find that

$$|r_{i+1}| \le G^i|r_1| + (I + G + \cdots + G^{i-1})g.$$

Writing $g := (\alpha I + \beta H)(|b| + |A||x|)$ and applying Lemma 1.2 we obtain

$$
\begin{aligned}
|(I + G + \cdots + G^{i-1})g| \le \big[\alpha + \big(&\alpha\|G\|_\infty(1 - \|G\|_\infty)^{-1} + \beta\|H\|_\infty(1 - \|G\|_\infty)^{-1}\big) \\
&\times \psi(|b| + |A||x|)\big](|b| + |A||x|).
\end{aligned}
$$

We summarize our findings in a theorem.

Theorem 4.1 *Let iterative refinement be applied to the nonsingular linear system $Ax = b$ of order n, using a solver satisfying* (2.2)*. There are matrices $M_i = I + O(u)$, $i = 1{:}2$, such that if*

$$G^i|r_1| \le \max(u, \overline{\gamma}_{n+1})(|b| + |A||x|) \qquad (4.10)$$

and

$$u + (1+u)\overline{\gamma}_{n+1} + (1 - \|G\|_\infty)^{-1}\big((u + (1+u)\overline{\gamma}_{n+1})\|G\|_\infty + (1+u)\overline{\gamma}_{n+1}\|H\|_\infty\big) \\ \times \psi(|b| + |A||x|) \le 2\max(u, \overline{\gamma}_{n+1}),$$

where

$$H = u(W + |A|)M_2|A^{-1}|, \qquad G = uI + (u + (1+u)\overline{\gamma}_{n+1})|A||A^{-1}| + HM_1,$$

then

$$\omega_{|A|,|b|}(\widehat{x}_{i+1}) \le \frac{3\max(u, \overline{\gamma}_{n+1})}{1 - 3\,\mathrm{cond}(A^{-1})\psi(|b| + |A||x|)\max(u, \overline{\gamma}_{n+1})}. \qquad \square$$

The gist of this result is that iterative refinement yields a small componentwise relative backward error provided that the solver is not too unstable ($\|W\|_\infty$ is not too large), A is not too ill conditioned ($\mathrm{cond}(A^{-1})$ is not too large), and the vector $|b| + |A||x|$ is not too badly scaled. The condition (4.10) is a necessary assumption that can fail to be satisfied for sufficiently large i only if $|b| + |A||x|$ has zero elements. Note that, roughly, $\|G\|_\infty \approx \max(u, \overline{\gamma}_{n+1})\|\,(W + |A|)|A^{-1}|\,\|_\infty$, and the residual is multiplied by a matrix of norm at most $\|G\|_\infty$ on each iteration.

Theorem 4.1 suggests that the only advantage of mixed precision iterative refinement over fixed precision iterative refinement for achieving a componentwise relative backward error of order u is that it tolerates a greater degree of instability, ill conditioning and bad scaling. The dependence of G on $\overline{u}$ is minor, as for the forward error analysis, so the theorem does not predict any significant difference in the rates of convergence of iterative refinement in mixed and fixed precision.

We turn our attention now to analyzing one step of fixed precision iterative refinement. From (4.9) and the inequality $|r_1| = |b - A\widehat{x}_1| \le uW|\widehat{x}_1|$, from (2.2), we have

$$|b - A\widehat{x}_2| \le uGW|\widehat{x}_1| + g. \qquad (4.11)$$

By considering the forms of G and g, we can glean a useful piece of insight immediately from (4.11): iterative refinement works because after just one step the matrix W occurring in the backward error bound for the solver is multiplied by u^2 in the residual bound; in other words, any instability in the solver is relegated to a second order term by the refinement process.

It is not possible to deduce a useful bound on $\omega_{|A|,|b|}(\widehat{x}_2)$ without making further assumptions on W. The most natural and useful assumption is that

$$W = Y|A|,$$

where, ideally, Y is of modest norm. Using this assumption we can derive a modified form of (4.11) that leads to a cleaner result. The trick is to bound $|A||\widehat{d}_i|$ directly. From (4.7) we have

$$|A||\widehat{d}_i| \le |A||A^{-1}|\big(M_1|r_i| + (1+u)\overline{\gamma}_{n+1}(|b| + |A||x|) + uY|A||\widehat{d}_i|\big),$$

or

$$(I - u|A||A^{-1}|Y)|A||\widehat{d}_i| \le |A||A^{-1}|\big(M_1|r_i| + (1+u)\overline{\gamma}_{n+1}(|b| + |A||x|)\big).$$

Hence, provided $u\|\,|A||A^{-1}|Y\,\|_\infty \le 1/2$, we have

$$|A||\widehat{d}_i| \le M_3|A||A^{-1}|\big(M_1|r_i| + (1+u)\overline{\gamma}_{n+1}(|b| + |A||x|)\big), \tag{4.12}$$

where $M_3 = (I - u|A||A^{-1}|Y)^{-1} \ge 0$ and $\|M_3\|_\infty \le 2$. The benefit of (4.12) is that it leads to a term $M_3|A||A^{-1}|$ in the bound instead of $|A|M_3|A^{-1}|$, and the norm of the former term is independent of the row scaling of A. Now, using (4.1) and (4.3) we have

$$|r_1| \le uY|A||\widehat{x}_1| \le uY(|A||x| + |A||A^{-1}||r_1|),$$

which implies

$$|r_1| \le uM_4Y|A||x|,$$

where $M_4 = (I - uY|A||A^{-1}|)^{-1}$ with $\|M_4\|_\infty \le 2$ if $\mathrm{cond}(A^{-1})\|Y\|_\infty u \le 1/2$. From (4.6) and (4.12) we have

$$\begin{aligned}|b - A\widehat{x}_2| &\le u(Y+I)M_3|A||A^{-1}|\big(uM_1M_4Y|A||x| + (1+u)\overline{\gamma}_{n+1}(|b| + |A||x|)\big)\\ &\qquad + uC_2M_4Y|A||x| + (1+u)\overline{\gamma}_{n+1}(|b| + |A||x|) + u|A||x|\\ &\le \big[u + (1+u)\overline{\gamma}_{n+1} + u(Y+I)M_3|A||A^{-1}|(uM_1M_4Y + (1+u)\overline{\gamma}_{n+1})\\ &\qquad + uC_2M_4Y\big](|b| + |A||x|).\end{aligned}$$

On invoking Lemma 1.2 and Corollary 1.3 we obtain the following result, which, in the special case where $\overline{u} = u^2$, is essentially Theorem 2.2 of Higham [4].

Theorem 4.2 *Let iterative refinement be applied to the nonsingular linear system $Ax = b$ of order n, using a solver satisfying (2.2) with $W = Y|A|$. There is a function*

$$f(u, \|Y\|_\infty) \approx q(u^2\|Y\|_\infty^2 + \max(u, \overline{\gamma}_{n+1})\|Y\|_\infty),$$

where q is a modest integer constant, such that if

$$\mathrm{cond}(A^{-1})f(u, \|Y\|_\infty)\psi(|b| + |A||x|) < 1$$

then

$$\omega_{|A|,|b|}(\widehat{x}_2) \le 3\max(u, \overline{\gamma}_{n+1}).$$

5 Practical Implications and LAPACK

We now discuss the implications of the results of the previous two sections for practical computation, with particular reference to LAPACK.

Mixed precision iterative refinement (MPIR) is relatively little used nowadays because it cannot be implemented in a portable manner in a double precision Fortran code. The main results are that as long as the solver is not too unstable, the matrix A is not too ill conditioned, and $|b| + |A||x|$ is not too badly scaled, MPIR yields a forward error of order u (Theorem 3.1) and a componentwise relative backward error of order u (Theorem 4.1). It is interesting to note that a componentwise relative backward error of order u does not imply a forward error of order u, but merely a forward error bound of order $\mathrm{cond}(A, x)u$, and neither does the reverse implication hold; therefore both the forward error analysis and the backward error analysis are needed.

Fixed precision iterative refinement (FPIR) is implemented in LAPACK by routines whose names end `-RFS`, which are called by the expert drivers (whose names end `-SVX`). FPIR is available in conjunction with LU-type factorizations for all the standard matrix types except triangular matrices, for which the original computed solution already has a componentwise relative backward error of order u. The `-RFS` routines terminate the refinement if the componentwise relative backward error $\omega = \omega_{|A|,|b|}(\widehat{x}_i)$ satisfies

1. $\omega \le u$,

2. ω has not decreased by a factor of at least 2 during the current iteration, or

3. five iterations have been performed.

These criteria were chosen to be robust in the face of different BLAS implementations and machine arithmetics. To justify the criteria we note that all the factorizations used in LAPACK are known to satisfy (2.2) with a reasonable bound on W (for proofs, see [7], for example). Theorem 4.1 therefore implies that FPIR will converge in all the `-RFS` routines provided A is not too ill conditioned and the vector $|b| + |A||x|$ is not too badly scaled. The second and third convergence criteria perform the practical necessity of terminating the refinement if convergence is not sufficiently fast. We mention that large or infinite values of $\psi(|b| + |A||x|)$ can occur when $a_{ij}x_j = 0$ for many i and j, as is most likely in sparse problems. Some possible ways to modify ω in the LAPACK stopping criterion in such situations are described by Arioli, Demmel and Duff [2].

An interesting question that remains is whether a single step of FPIR guarantees that $\omega_{|A|,|b|}(\widehat{x}_i) \approx u$. Theorem 4.2 gives a positive answer for solvers for which $W = Y|A|$ with a modestly normed Y, with the usual provisos that A is not too ill conditioned and the vector $|b| + |A||x|$ is not too badly scaled. Such solvers include those based on a QR factorization computed by Householder transformations, Givens rotations, or the modified Gram-Schmidt method (see [7, Ch. 18]). For an LU factorization with computed LU factors $\widehat{L}$ and $\widehat{U}$ we have, from (3.5),

$$uW \equiv 2\gamma_n|\widehat{L}||\widehat{U}| \approx 2\gamma_n|\widehat{L}||\widehat{L}^{-1}A| \leq uY|A|, \qquad Y \approx 2n|\widehat{L}||\widehat{L}^{-1}|. \tag{5.1}$$

Without pivoting, $\|Y\|_\infty$ can be arbitrarily large. With partial pivoting we have $|l_{ij}| \leq 1$, and although $\|\,|\widehat{L}||\widehat{L}^{-1}|\,\|_\infty$ can be as large as 2^n-1, it is typically of order n in practice [13]. We can summarize by saying that for Gaussian elimination with partial pivoting one step of FPIR will usually be enough to yield a small componentwise relative backward error as long as A is not too ill conditioned and $|b| + |A||x|$ is not too badly scaled, which, of course, is Skeel's main result from [12].

The other two factorizations for which LAPACK supports FPIR are Cholesky factorization and the block $\mathrm{LDL}^{\mathrm{T}}$ factorization computed by the diagonal pivoting method. For the Cholesky factorization $A = R^TR$, where R is upper triangular with positive diagonal elements, we can take $uW = 2\gamma_{n+1}|\widehat{R}^T||\widehat{R}|$ in (2.2) [7, Th. 10.4]. If we bound $W \leq Y|A|$ using the same approach as in (5.1) we find that $Y \approx 2n(|\widehat{R}^{-1}||\widehat{R}|)^T$, which is unbounded. However, for the Cholesky factorization with complete pivoting, $\Pi^TA\Pi = R^TR$, the pivoting causes R to satisfy inequalities that imply $\|\,|R^{-1}||R|\,\|_\infty \leq 2^n - 1$ [7, Lem. 8.6], so we have a similar result as for GEPP. Interestingly, our practical experience is that complete pivoting in Cholesky factorization has little effect on the performance of iterative refinement.

The diagonal pivoting method computes a factorization $PAP^T = LDL^T$, where L is unit lower triangular, D is block diagonal with 1×1 and 2×2 diagonal blocks, and P is a permutation matrix. LAPACK uses the partial pivoting strategy of Bunch and Kaufman [3], for which the backward error result (2.2) holds with

$$uW = p(n)u\big(|A| + P^T|\widehat{L}||\widehat{D}||\widehat{L}^T|P\big) + O(u^2),$$

and, furthermore, $\|uW\|_\infty \leq p(n)u\rho_n\|A\|_\infty$, where p is a quadratic and ρ is the growth factor; see Higham [5]. Attempting to bound $W \leq Y|A|$ using the approach in (5.1) does not give a useful bound for $\|Y\|_\infty$.

In conclusion, we are not able to guarantee that "one step of FPIR is enough" for Cholesky factorization or for the diagonal pivoting method, but LAPACK's use of FPIR with these factorizations is, nevertheless, justified by Theorem 4.1.

6 Numerical Experiments

We give two numerical examples to illustrate some of the salient features of mixed and fixed precision iterative refinement. The computations were performed in MATLAB, using simulated IEEE single precision arithmetic in which we rounded the result of every arithmetic operation to 24 significant bits; therefore $u = 2^{-24} \approx 5.96 \times 10^{-8}$. To implement MPIR we simply computed residuals using MATLAB's double precision arithmetic.

The first example is for Gaussian elimination (GE) without pivoting, applied to the scaled 15×15 orthogonal matrix A_1 with $a_{ij} = d_i(2/(n+1))^{1/2}\sin(ij\pi/(n+1))$, where $d(1\colon n) = \alpha^i$, with $\alpha^n = 10^{-5}$. (This matrix is a row-scaled version of `orthog(15)` from the Test Matrix Toolbox [6], which is the eigenvector matrix for the second difference matrix.) The right-hand side b is generated as $b = A[1,2,\dots,15]^T$.

The second example applies Gaussian elimination with partial pivoting (GEPP) to a random 10×10 matrix A_2 with $\kappa_2(A) = 10^6$. (This matrix is generated as `randn('seed',1); A = randsvd(10, 1e6)`, using the Test Matrix Toolbox [6].) The right-hand side is selected as in the first example.

The results are shown in Tables 6.1–6.2. For the matrix W we take $2nP^T|\widehat{L}||\widehat{U}|$, where $PA \approx \widehat{L}\widehat{U}$ is the computed LU factorization ($P = I$ for GE). We make several observations.

1. In the first example, GE yields a moderately large componentwise relative backward error, partly because the growth factor $\rho_{15} \approx 3112$. FPIR achieves $\omega_{|A|,|b|} \le u$ after 3 iterations, even though the product $\mathrm{cond}(A^{-1})\psi(|b| + |A||x|)$ exceeds u^{-1}, so that the conditions in Theorem 4.1 are not satisfied. This is a common occurrence: iterative refinement often works well even for problems that are so extreme that the analysis does not guarantee success. Since $\mathrm{cond}(A,x)$ is of order 1, the forward error matches the behaviour of the componentwise relative backward error. Note that MPIR is no more effective than FPIR at achieving $\omega_{|A|,|b|} \le u$, though the $\omega_{|A|,|b|}$ values do converge for MPIR, unlike for FPIR.

2. The first example emphasizes how FPIR can overcome the effects of poor scaling. The standard condition number $\kappa_\infty(A)$ is of order 10^5 due to the bad row scaling of A, while $\mathrm{cond}(A)$ and $\mathrm{cond}(A,x)$ are of order 1. FPIR produces a solution with forward error of order u, as we would hope in view of Theorem 3.2, even though the theorem is not strictly applicable since $\eta > 1$. (If we use GEPP instead of GE in the first example, the behaviour is broadly the same.)

3. For the second example, GEPP achieves a componentwise relative backward error of order u, so FPIR is not worthwhile. MPIR is beneficial, however: it reduces the forward error to order u, as predicted by Theorem 3.1. This example shows how the convergence test must be chosen to reflect the desired benefits of iterative refinement, for if the iteration were terminated when $\omega_{|A|,|b|} \le u$ then MPIR would not be performed at all.

7 Concluding Remarks

The analysis we have presented is sufficiently general to cover all existing applications of iterative refinement for linear systems—in mixed or fixed precision, with one or

Table 6.1: Result for GE with orthogonal matrix A_1.

$$\mathrm{cond}(A) = 1.26e1,\ \mathrm{cond}(A,x) = 6.72e0,\ \kappa_\infty(A) = 1.81e5$$
$$\mathrm{cond}(A^{-1}) = 1.65e5,\ \psi(|b| + |A||x|) = 1.98e5$$
$$u\|\,|A^{-1}|(|A| + W)\,\|_\infty = 3.89e0.$$

FPIR: Iteration	$\omega_{\lvert A\rvert,\lvert b\rvert}(\widehat{x}_i)$	$\Vert x - \widehat{x}_i\Vert_\infty / \Vert x\Vert_\infty$
0	9.85e-3	1.34e-2
1	4.04e-5	4.38e-5
2	5.16e-8	9.75e-8
3	1.43e-8	2.35e-8
4	1.54e-8	3.75e-8
5	2.78e-8	7.02e-8
6	1.65e-8	4.72e-8
7	2.52e-8	6.14e-8
8	1.55e-8	2.74e-8
9	2.31e-8	4.68e-8

MPIR: Iteration	$\omega_{\lvert A\rvert,\lvert b\rvert}(\widehat{x}_i)$	$\Vert x - \widehat{x}_i\Vert_\infty / \Vert x\Vert_\infty$
0	9.85e-3	1.34e-2
1	3.26e-5	3.91e-5
2	1.05e-7	1.39e-7
3	1.06e-8	2.35e-8
4	1.06e-8	2.35e-8
5	1.06e-8	2.35e-8

Table 6.2: Result for GEPP with random, ill conditioned matrix A_2.

$$\mathrm{cond}(A) = 1.04e6,\ \mathrm{cond}(A,x) = 5.62e5,\ \kappa_\infty(A) = 2.38e6$$
$$\mathrm{cond}(A^{-1}) = 1.30e6,\ \psi(|b| + |A||x|) = 8.29e0$$
$$u\|\,|A^{-1}|(|A| + W)\,\|_\infty = 3.86\text{e-}1.$$

FPIR: Iteration	$\omega_{\lvert A\rvert,\lvert b\rvert}(\widehat{x}_i)$	$\|x - \widehat{x}_i\|_\infty / \|x\|_\infty$
0	2.34e-8	2.79e-3
1	3.66e-8	3.59e-3
2	2.11e-8	1.33e-3
3	4.71e-8	9.38e-3
4	3.95e-8	2.49e-3
5	2.49e-8	3.07e-3
6	3.42e-8	4.12e-3
7	1.88e-8	2.33e-3
8	2.81e-8	7.17e-4
9	2.45e-8	5.27e-3

MPIR: Iteration	$\omega_{\lvert A\rvert,\lvert b\rvert}(\widehat{x}_i)$	$\|x - \widehat{x}_i\|_\infty / \|x\|_\infty$
0	2.34e-8	2.79e-3
1	1.73e-8	1.49e-5
2	1.94e-8	7.37e-8
3	2.30e-8	2.85e-8
4	2.30e-8	2.85e-8

more iterations—and it fully supports the use of iterative refinement in LAPACK. One interesting question remains: is one step of fixed precision iterative refinement enough for Cholesky factorization to produce a small componentwise relative backward error? It seems to be generally true that any result for LU factorization has an analogue for Cholesky factorization that is at least as strong, yet our analysis does not give a "one step is enough" result for Cholesky factorization. A scaling argument can be used to replace A in the bounds by $H = D^{-1}AD$, where $h_{ii} = 1$, and a result of Van der Sluis [14] implies that $\kappa_2(H) \leq n \min\{\, \kappa_2(FAF) : F \text{ diagonal} \,\}$; however, $\kappa_2(H)$ can still be large and the scaling changes the term $\psi(|b| + |A||x|)$. Therefore we pose the open problem: prove that "one step is enough" for Cholesky factorization, or find a numerical counterexample (with $\operatorname{cond}(A^{-1}) f(u, \|Y\|_\infty) \psi(|b| + |A||x|)$ sufficiently less than 1, in the notation of Theorem 4.2). The corresponding problem for the factorization produced by the diagonal pivoting method is also open.

References

[1] E. Anderson, Z. Bai, C. H. Bischof, J. W. Demmel, J. J. Dongarra, J. J. Du Croz, A. Greenbaum, S. J. Hammarling, A. McKenney, S. Ostrouchov, and D. C. Sorensen. *LAPACK Users' Guide, Release 2.0.* Second edition, Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 1995. xix+325 pp. ISBN 0-89871-345-5.

[2] M. Arioli, J. W. Demmel, and I. S. Duff. Solving sparse linear systems with sparse backward error. *SIAM J. Matrix Anal. Appl.*, 10(2):165–190, 1989.

[3] James R. Bunch and Linda Kaufman. Some stable methods for calculating inertia and solving symmetric linear systems. *Math. Comp.*, 31(137):163–179, 1977.

[4] Nicholas J. Higham. Iterative refinement enhances the stability of QR factorization methods for solving linear equations. *BIT*, 31:447–468, 1991.

[5] Nicholas J. Higham. Stability of the diagonal pivoting method with partial pivoting. Numerical Analysis Report No. 265, Manchester Centre for Computational Mathematics, Manchester, England, July 1995. 17 pp.

[6] Nicholas J. Higham. The Test Matrix Toolbox for MATLAB (version 3.0). Numerical Analysis Report No. 276, Manchester Centre for Computational Mathematics, Manchester, England, September 1995. 70 pp.

[7] Nicholas J. Higham. *Accuracy and Stability of Numerical Algorithms.* Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 1996. Approx xxiv+690 pp. In press. ISBN 0-89871-355-2.

[8] Nicholas J. Higham and Philip A. Knight. Componentwise error analysis for stationary iterative methods. In *Linear Algebra, Markov Chains, and Queueing Models*, Carl D. Meyer and Robert J. Plemmons, editors, volume 48 of *IMA Volumes in Mathematics and its Applications*, Springer-Verlag, New York, 1993, pages 29–46.

[9] M. Jankowski and H. Woźniakowski. Iterative refinement implies numerical stability. *BIT*, 17:303–311, 1977.

[10] Cleve B. Moler. Iterative refinement in floating point. *J. Assoc. Comput. Mach.*, 14(2):316–321, 1967.

[11] W. Oettli and W. Prager. Compatibility of approximate solution of linear equations with given error bounds for coefficients and right-hand sides. *Numer. Math.*, 6:405–409, 1964.

[12] Robert D. Skeel. Iterative refinement implies numerical stability for Gaussian elimination. *Math. Comp.*, 35(151):817–832, 1980.

[13] Lloyd N. Trefethen and Robert S. Schreiber. Average-case stability of Gaussian elimination. *SIAM J. Matrix Anal. Appl.*, 11(3):335–360, 1990.

[14] A. van der Sluis. Condition numbers and equilibration of matrices. *Numer. Math.*, 14:14–23, 1969.

[15] J. H. Wilkinson. Progress report on the Automatic Computing Engine. Report MA/17/1024, Mathematics Division, Department of Scientific and Industrial Research, National Physical Laboratory, Teddington, UK, April 1948. 127 pp.

[16] J. H. Wilkinson. *Rounding Errors in Algebraic Processes.* Notes on Applied Science No. 32, Her Majesty's Stationery Office, London, 1963. vi+161 pp. Also published by Prentice-Hall, Englewood Cliffs, NJ, USA. Reprinted by Dover, New York, 1994. ISBN 0-486-67999-3.