IMA Journal of Numerical Analysis (2007) **27**, 219–231
doi:10.1093/imanum/drl037
Advance Access publication on November 29, 2006

A model for understanding numerical stability

FOLKMAR BORNEMANN†
Zentrum Mathematik, Technische Universität München, Boltzmannstrasse 3, 85747 Garching, Germany

[Received on 20 March 2006; revised on 24 October 2006]

We present a model of roundoff error analysis that combines simplicity with predictive power. Though not considering all sources of roundoff within an algorithm, the model is related to a recursive roundoff error analysis and therefore is capable of correctly predicting stability or instability of an algorithm. By means of nontrivial examples, such as the componentwise backward stability analysis of Gaussian elimination with a single iterative refinement step, we demonstrate that the model even yields quantitative backward error bounds that show all the known problem-dependent terms with the exception of dimension-dependent constants. The model can serve as a convenient tool for teaching or as a heuristic device to discover stability results before entering a further detailed analysis.

Keywords: numerical stability; model of roundoff error analysis; Gaussian elimination.

1. Introduction

An algorithm for the numerical evaluation of a complicated function f is just a decomposition into simple intermediate steps, such as arithmetic operations, elementary transcendental functions or well-behaved and well-understood library algorithms (e.g. matrix multiplication):

$$f = g_1 \circ g_2 \circ \cdots \circ g_k.$$

In floating-point arithmetic, each of these intermediate steps is contaminated by roundoff and hence contributes to the final perturbation of the result in a two-fold fashion: first, by 'generating' roundoff error itself and second, by 'propagating' the roundoff errors of previous steps. Since the early days of numerical computing, there has been much progress in clarifying the underlying structure and organizing the results in a concise, easily interpreted form. However, a detailed analysis (Higham, 2002) is still often quite involved and remains a battlefield for experts, too tedious to teach and explain in detail beyond the most trivial cases in a beginner's course on numerical analysis. The instructor typically chooses between two options: skipping the nontrivial results (such as stability of Gaussian elimination) entirely or just stating the results without proofs. Either choice is unsatisfactory for good students since they cannot develop an 'understanding' of the mathematical structure and reasons.

We will demonstrate in this paper that the overall behaviour of an algorithm can very often be well understood by analysing a simplified model of the sources of roundoff error. As in the natural sciences, such a model has to balance simplicity with predictive ability. If such a simple model leads to the same predictions, qualitatively and perhaps even quantitatively, as a full-fledged *a priori* roundoff error analysis, we may rightly claim to have contributed to the understanding of the algorithm's behaviour.

†Email: bornemann@ma.tum.de

© The author 2006. Published by Oxford University Press on behalf of the Institute of Mathematics and its Applications. All rights reserved.

In fact, all the estimates of our model analysis that we present in this paper will give the 'same' estimates as a detailed *a priori* analysis—with the exception of the dimension-dependent constants, which are, however, anyway the least important part of any roundoff error analysis (Higham, 2002, p. 65). In particular, with just a few lines of simple calculations, we obtain the nontrivial results on the norm- and componentwise backward stability of Gaussian elimination ranging from the early work (Wilkinson, 1963) to the analysis of iterative refinement (Skeel, 1980).

In addition to being a convenient (and, to the experience of the author, also a successful) tool in teaching, our model might serve as a 'heuristic' device in discovering the structure of a stability result—before one enters, in a second step, taking advantage of the obtained knowledge, a fully detailed roundoff error analysis.

1.1 *The model*

The roundoff error analysis that we propose is based on the observation that in many if not most cases, a critical intermediate step can be identified that leads to a natural decomposition

$$f = \underbrace{g_1 \circ \cdots \circ g_j}_{=h} \circ \underbrace{g_{j+1} \circ \cdots \circ g_k}_{=g} = h \circ g$$

into just two fundamental steps. Now, the model is based on the 'simplifying assumption' that roundoff error just affects the single intermediate result—after being output by g, before being input to h. That is, we analyse the error of the 'realization map'

$$\tilde{f} = h \circ \mathrm{fl} \circ g.$$

Here, $\mathrm{fl}\colon \mathbb{R}^p \to \mathbb{G}^p$ denotes the componentwise rounding, subject to the standard model of floating-point arithmetic

$$|\mathrm{fl}(x) - x| \leqslant u \cdot |x|,$$

where u denotes the unit roundoff of the arithmetic ($u \approx 1.11 \times 10^{-16}$ for IEEE double precision) and $\mathbb{G}$ the floating-point numbers. We understand $|x|$ to be componentwise for vectors and matrices.

1.2 *Outline of the paper*

In Section 2, we analyse the backward stability of the realization map $\tilde{f}$, which turns out to be determined by the condition number of g^{-1}. We will specify the relation of the model to a complete analysis. In fact, if the model is unstable, the same has to be expected for the real situation. On the other hand, if g and h are realized by the backward stable algorithms, then the resulting algorithm for f would inherit the stability of the model. This helps to understand the success of our model and suggests a recursive approach to a full roundoff error analysis.

The rest of the paper studies some algorithms for the solution of a linear system $Ax = b$. In Section 3, we recall some classic expressions for the backward error of linear systems that are the points of departure for the simple estimates to follow. In Section 4, we study the naive algorithm, i.e. multiplication with A^{-1}, and show its instability for badly conditioned matrices. In Section 5, we study the normwise backward error of Gaussian elimination and obtain the classic result (Wilkinson, 1963). In Section 6, we get the result (Skeel, 1979) on the componentwise backward error of Gaussian elimination, correctly predicting the influence of the scaling of the system. Finally, in Section 7, we show how to discover within the framework of our model the result (Skeel, 1980) that a single step of iterative refinement implies componentwise backward stability of Gaussian elimination.

2. Backward stability

The main result of a 'qualitative' study of our model can be summarized as follows:

> Backward stability requires that g^{-1} be well conditioned.

In fact, backward stability analysis requires the result of the algorithm for an input x, i.e. $\tilde{f}(x)$ in our model, to be represented as the 'exact' solution to perturbed data: $\tilde{f}(x) = f(x + \Delta x)$. Writing $w = g(x)$ for short, we have

$$\tilde{f}(x) = h(\mathrm{fl}(w)) = h(w + \Delta w), \quad |\Delta w| = |\mathrm{fl}(w) - w| \leqslant u \cdot |w|.$$

Assuming g to be invertible, we propagate Δw backwards to obtain an estimate for Δx:[1]

$$x + \Delta x = g^{-1}(w + \Delta w), \quad |\Delta x| \leqslant \kappa_{g^{-1}} u \cdot |x| + \mathrm{O}(u^2),$$

where the smallest constant $\kappa_{g^{-1}}$ defines the (componentwise) relative condition number of g^{-1} at w. Hence, the backward error is bounded by the unit roundoff amplified by $\kappa_{g^{-1}}$.

2.1 *Examples*

A. Consider the evaluation of $f(x) = \log^2(1+x)$ for $x \approx 0$. A direct implementation of the defining formula corresponds to the decomposition

$$f\colon x \overset{g}{\longmapsto} w = 1 + x \overset{h}{\longmapsto} \log^2(w).$$

Now, because $w = 1 + x \approx 1$, the inverse function $g^{-1}\colon w \mapsto x = w - 1$ is a subtraction in the cancellation regime, thus badly conditioned. Hence, we predict the instability of the formula, which simple examples confirm. In fact, here the bad conditioning of g^{-1} reflects the 'loss of information' in g. We have $\mathrm{fl}(g(x)) = \mathrm{fl}(1 + x) = 1$ as soon as x is smaller than the resolution of the machine arithmetic. In general, well conditioning of g^{-1}, however, requires that the input x is accurately reconstructable from the intermediate result $w = g(x)$.

B. The solution $x \in \mathbb{R}^m$ of a linear system of equations $Ax = b$ with a nonsingular $A \in \mathbb{R}^{m\times m}$ can formally be represented as $x = A^{-1} \cdot b$. This suggests the naive algorithm corresponding to the decomposition

$$f\colon A \overset{g}{\longmapsto} A^{-1} \overset{h}{\longmapsto} x = A^{-1} \cdot b.$$

Now, $g^{-1}\colon A^{-1} \mapsto A$ is just g again; its condition is (in the normwise case) the condition number of the matrix A. Thus, we expect the algorithm to be unstable for certain badly conditioned matrices. Examples that display such instability will be given in Section 4, where we extend our analysis to a more quantitative setting.

C. On the other hand, the solution of the linear system $Ax = b$ by Gaussian elimination corresponds to the decomposition

$$f\colon A \overset{g}{\longmapsto} (L, U) \overset{h}{\longmapsto} x.$$

[1] On the other hand, if g is many-to-one, we choose Δx in such a way that the corresponding constant $\kappa_{g^{-1}}$ is as small as possible. Again, this constant will be called the relative condition number of g^{-1}. In practice, one often uses a specific selection of Δx and works with the corresponding upper bound of $\kappa_{g^{-1}}$. If there is no such Δx, we simply put $\kappa_{g^{-1}} = \infty$.

Here, g represents the LU factorization step, whereas h represents the substitution steps. Now, the inverse of g, i.e.

$$g^{-1}\colon (L, U) \longmapsto A = L \cdot U,$$

is just a 'matrix multiplication'. Its condition number can be estimated by

$$\kappa_{g^{-1}} \leqslant 2\frac{\||L|\cdot|U|\|}{\|A\|},$$

which is, as will be discussed in more detail in Section 5, sufficient to explain the instabilities to be observed for Gaussian elimination with or without partial pivoting.

2.2 *Relation of the model to a complete analysis*

In fact, the condition number of g^{-1} turns out to be relevant for a full roundoff error analysis, too. Here, we could recursively define the realization of $f = h \circ g$ by

$$\tilde{f} = \tilde{h} \circ \tilde{g},$$

starting with the backward stable realization of the arithmetic operations and the basic elementary functions. (Of course, in general we cannot assume that in each step of this recursion the g-part of the decomposition is invertible. However, it is possible to give a reasonably simple definition of $\kappa_{g^{-1}}$, even if g is many-to-one; see footnote 1.)

With $[\![\Delta x]\!]$ denoting the maximum componentwise relative error,[2] we define the smallest number $\beta_f \geqslant 0$ such that

$$\tilde{f}(x) = f(x + \Delta x), \quad [\![\Delta x]\!] \leqslant \beta_f \cdot u + \mathrm{O}(u^2)$$

as the 'stability indicator' of $\tilde{f}$. Backward stability requires β_f to be not too large.

LEMMA 2.1 For g invertible, there holds the recursive estimate

$$\beta_f \leqslant \beta_g + \kappa_{g^{-1}} \cdot \beta_h. \tag{2.1}$$

Proof. The stability indicator of $\tilde{h}$ gives

$$\tilde{f}(x) = \tilde{h}(\tilde{g}(x)) = h(\tilde{g}(x) + \Delta w), \quad [\![\Delta w]\!] \leqslant \beta_h \cdot u + \mathrm{O}(u^2).$$

The stability indicator of $\tilde{g}$ and the relative condition number of g^{-1} allow for the estimates

$$\begin{aligned}\tilde{g}(x) + \Delta w &= g(x + \Delta x_1) + \Delta w, \quad [\![\Delta x_1]\!] \leqslant \beta_g \cdot u + \mathrm{O}(u^2),\\ &= g(x + \Delta x_1 + \Delta x_2), \quad [\![\Delta x_2]\!] \leqslant \kappa_{g^{-1}} \cdot [\![\Delta w]\!] + \mathrm{O}([\![\Delta w]\!]^2).\end{aligned}$$

Since Δx_1 and Δx_2 are both perturbations of the same quantity x, there holds the triangle inequality for relative errors

$$\Delta x = \Delta x_1 + \Delta x_2, \quad [\![\Delta x]\!] \leqslant [\![\Delta x_1]\!] + [\![\Delta x_2]\!] \leqslant (\beta_g + \kappa_{g^{-1}} \cdot \beta_h)u + \mathrm{O}(u^2),$$

and we get the assertion. □

[2] That is, for the perturbation $\Delta x \in \mathbb{R}^m$ of a quantity $x \in \mathbb{R}^m$, we have $[\![\Delta x]\!] = \max_{j=1:m} |\Delta x_j|/|x_j|$ with the convention that $0/0 = 0$.

Thus, we may complement the maxim from the beginning of this section by the following rule:

> If g^{-1} is well-conditioned, backward stable realizations of g and h induce a backward stable realization of f.

Summarizing, the logical status of the proposed model is as follows: If the model predicts instability, we can expect instability in reality—independently of how g and h are realized in practice. Most probably, in examples that realize the worst case scenario of the condition number bound, there will be instability even if g and h were calculated exactly, a fact which certainly shakes our faith in the algorithm. On the other hand, if the model predicts stability, the actual stability of the algorithm depends on how g and h are realized algorithmically. In the framework of backward stability, stability of the realization of g and h implies stability of the resulting algorithm for f.

EXAMPLE. Let us illustrate these points by reconsidering the example A of Section 2.1. Here, we decompose $f(x) = \log^2(1+x)$, $x \approx 0$, differently into

$$f\colon x \overset{g}{\longmapsto} w = \log(1+x) \overset{h}{\longmapsto} w^2.$$

Now, the critical map $g^{-1}\colon w \mapsto \mathrm{e}^w - 1$ has the relative condition number $\kappa_{g^{-1}} \approx 1$ for $w \approx 0$. The model alone would therefore predict numerical stability. On the other hand, the full, recursive analysis has to take the actual algorithms for g and h into account. Step h, as a multiplication in IEEE arithmetic, is certainly backward stable. However, the status of g is far less clear. If its realization is chosen to be based on the decomposition $g\colon x \mapsto z = 1 + x \mapsto \log(z)$, then an analysis similar to the example A of Section 2.1 reveals instability. On the other hand, if g is realized, for instance, by using Kahan's stable algorithm as implemented in Matlab's log1p command, then the resulting algorithm for f is stable, too.

Hence, the choice of the decomposition will critically determine the success or failure of the model. In general, making a conclusive choice will depend on the user's experience or luck. However, we will show in the rest of the paper that quite naturally such decompositions occur in the analysis of the stability of Gaussian elimination.

3. The backward error of linear systems

To prepare for a more quantitative analysis of algorithms for the solution of linear systems of equations $Ax = b$, we recall the concept of the backward error of an output vector $\tilde{x} \in \mathbb{R}^m$. 'Normwise' analysis considers[3]

$$\eta = \min_{E \in \mathbb{R}^{m\times m}} \left\{ \frac{\|E\|}{\|A\|} \colon (A+E)\tilde{x} = b \right\},$$

whereas 'componentwise' analysis studies

$$\omega = \min_{E \in \mathbb{R}^{m\times m}} \left\{ \max_{ij} \frac{|E|_{ij}}{|A|_{ij}} \colon (A+E)\tilde{x} = b \right\}.$$

The classic results of Rigal & Gaches (1967) and Oettli & Prager (1964) show that η and ω can be calculated from the data of the linear system and the output vector $\tilde{x}$ by means of the following simple

[3]Throughout the paper, we deal with 'monotone' vector norms like the 1-, 2- or ∞-norm and the induced matrix norms.

formulae:

$$\eta = \frac{\|r\|}{\|A\|\cdot\|\tilde{x}\|}, \quad \omega = \max_{j=1:m} \frac{|r_j|}{(|A|\cdot|\tilde{x}|)_j}, \tag{3.1}$$

where $r = b - A\tilde{x}$ denotes the 'residual' of $\tilde{x}$. These formulae, which have very short and straightforward proofs (Higham, 2002, pp. 120/122), are also valuable for the *a posteriori* assessment of computed solutions. We will use them as a convenient point of departure for a quantitative analysis in the frame of our proposed model.

4. Model analysis of the naive algorithm for linear systems

As discussed in Section 2.1, the naive algorithm for the solution of a linear system is given by the decomposition

$$f\colon A \overset{g}{\longmapsto} B = A^{-1} \overset{h}{\longmapsto} x = B \cdot b.$$

Our model analyses how roundoff in B affects the solution x and its backward error:

$$\tilde{f}\colon A \overset{g}{\longmapsto} B = A^{-1} \overset{\mathrm{fl}}{\longmapsto} \tilde{B} = B + \Delta B \overset{h}{\longmapsto} \tilde{x} = \tilde{B} \cdot b.$$

The perturbation $|\Delta B| \leqslant u \cdot |B|$ induces, by propagating backwards through g^{-1}, an equivalent perturbation $\tilde{A} = A + \Delta A = g^{-1}(\tilde{B})$ of the input matrix. By construction, we have $\tilde{A}\tilde{x} = b$,

$$(A + \Delta A)(A^{-1} + \Delta B) = I, \quad \text{i.e. } \Delta A = -A \cdot \Delta B \cdot A - \Delta A \cdot \Delta B \cdot A,$$

and therefore the componentwise estimate

$$|\Delta A \cdot \tilde{x}| \leqslant |A|\cdot|A^{-1}|\cdot|A\tilde{x}|u + \mathrm{O}(u^2).$$

Since $r = b - A\tilde{x} = \Delta A\cdot\tilde{x}$ and $\tilde{x} = x + \mathrm{O}(u)$, we get by (3.1)

$$\eta = \frac{\|\Delta A \cdot \tilde{x}\|}{\|A\|\cdot\|\tilde{x}\|} \leqslant \frac{\||A|\cdot|A^{-1}|\cdot|Ax|\|}{\|A\|\cdot\|x\|}u + \mathrm{O}(u^2) =: \gamma(A,x)u + \mathrm{O}(u^2). \tag{4.1}$$

To relate with better known quantities, we may further estimate

$$\gamma(A,x) \leqslant \||A|\cdot|A^{-1}|\| = \mathrm{cond}(A^{-1}),$$

in agreement with our qualitative analysis of the example B of Section 2.1. Thus, instability (in the sense of large normwise backward errors) appears to be possible only for badly conditioned matrices.

4.1 *Examples*[4]

A. A notoriously badly conditioned matrix is the famous 'Hilbert matrix' H_m for larger dimensions m. In Matlab, there is a command 'invhilb' that supplies H_m^{-1} and allows us to implement the

[4] If not explicitly stated otherwise, all the examples in this paper use the norm $\|\cdot\|_\infty$.

naive algorithm:[5]

```
>> m = 20; A = hilb(m); B = invhilb(m); b = ones(m,1); x = B*b;
>> eta = norm(b - A*x,inf)/norm(A,inf)/norm(x,inf)

eta = 1.2787e-005
```

Thus, the naive algorithm is unstable as predicted by the *a priori* bound (4.1), which turns out to be

$$\eta = 1.27\cdots \times 10^{-5} \leqslant \gamma(A, x) \cdot u = 5.69\cdots \times 10^{-4},$$

a fairly good prediction indeed. On the other hand, we have to be careful to base a prediction on coarser upper bounds that were introduced for the ease of interpretation: the condition number yields

$$\eta \leqslant \mathrm{cond}(A^{-1}) \cdot u = 6.63\cdots \times 10^{11},$$

which gives too pessimistic a picture of the actual backward error.

B. The following example (Skeel, 1979, p. 509) shows that the naive algorithm can be stable for 'some' badly conditioned matrices:

$$A = \begin{pmatrix} 1 & 1 & -1 & -1 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & \epsilon & 0 \\ 0 & \epsilon & 0 & 0 \end{pmatrix}, \quad b = \begin{pmatrix} 0 \\ 2 \\ 1 \\ 1 \end{pmatrix}, \quad x = \begin{pmatrix} 1 \\ \epsilon^{-1} \\ \epsilon^{-1} \\ 1 \end{pmatrix}.$$

This matrix A satisfies

$$\mathrm{cond}(A) = 4, \quad \mathrm{cond}(A^{-1}) = 2 + 4\epsilon^{-1}.$$

However, numerical experiments with various small $0 < \epsilon \ll 1$ exhibit very small backward errors of about the size of the unit roundoff. This is fully reflected by our model analysis, since

$$\gamma(A, x) = 1 + \frac{\epsilon}{2} \approx 1.$$

5. Model analysis of Gaussian elimination: the normwise case

As discussed in the example C of Section 2.1, the solution of a linear system $Ax = b$ by Gaussian elimination corresponds to the decomposition

$$f\colon A \overset{g}{\longmapsto} (L, U) \overset{h}{\longmapsto} x.$$

In the model, the roundoff affects only the intermediate result, the LU factorization, by

$$\tilde{f}\colon A \overset{g}{\longmapsto} (L, U) \overset{\mathrm{fl}}{\longmapsto} (\tilde{L}, \tilde{U}) = (L + \Delta L, U + \Delta U) \overset{h}{\longmapsto} \tilde{x}.$$

[5]Here and in the examples to follow, we have cross-checked the 'actually calculated' backward errors with higher precision arithmetic. The first digits were always correct, so that the conclusions we draw are not affected by roundoff errors in the computed residuals.

Here, the perturbations $|\Delta L| \leqslant u \cdot |L|$ and $|\Delta U| \leqslant u \cdot |U|$ induce, by propagating through the inverse of g (i.e. matrix multiplication), an equivalent perturbation of the input matrix

$$A + \Delta A = \tilde{L}\tilde{U} = (L + \Delta L) \cdot (U + \Delta U), \quad \text{i.e. } \Delta A = \Delta L \cdot U + L \cdot \Delta U + \Delta L \cdot \Delta U.$$

This way, we obtain the componentwise estimate

$$|\Delta A| \leqslant 2|L||U| \cdot u_*, \quad u_* = u + u^2/2. \tag{5.1}$$

Because of $r = b - A\tilde{x} = \Delta A \cdot \tilde{x}$, we get by (3.1)

$$\eta \leqslant \frac{\||\Delta A|\cdot|\tilde{x}|\|}{\|A\|\cdot\|\tilde{x}\|} \leqslant 2\frac{\||L|\cdot|U|\cdot|\tilde{x}|\|}{\|A\|\cdot\|\tilde{x}\|}u_* \leqslant 2\frac{\||L|\cdot|U|\|}{\|A\|}u_* =: 2\gamma(L,U)u_*, \tag{5.2}$$

in agreement with our qualitative analysis of the example C of Section 2.1. If we restrict ourselves to monotone matrix norms, we can further estimate the 'growth factor' $\gamma(L,U)$ by using $U = L^{-1} \cdot A$

$$\gamma(L,U) = \frac{\||L|\cdot|U|\|}{\|A\|} \leqslant \frac{\||L|\cdot|L^{-1}|\|\cdot\|A\|}{\|A\|} = \mathrm{cond}(L^{-1}).$$

Thus, an instability of Gaussian elimination in the normwise case requires a badly conditioned L factor of the matrix A.

5.1 *Examples*

A. It is well-known that Gaussian elimination without pivoting is bound to be 'unstable' for small pivot elements. An example is given by

$$A = \begin{pmatrix} \epsilon & 1 \\ 1 & 1 \end{pmatrix}, \quad L = \begin{pmatrix} 1 & 0 \\ \epsilon^{-1} & 1 \end{pmatrix}, \quad U = \begin{pmatrix} \epsilon & 1 \\ 0 & 1-\epsilon^{-1} \end{pmatrix}.$$

For $\epsilon = u$, $b = (1,0)^{\mathrm{T}}$, a numerical experiment yields[6] $\tilde{x} \doteq (-2,1)$; the exact solution, however, would be $x \doteq (-1,1)^{\mathrm{T}}$. The backward error turns out to be $\eta \doteq \frac{1}{4}$. On the other hand, we have

$$\gamma(L,U) = \epsilon^{-1}, \quad \mathrm{cond}(L^{-1}) = 1 + 2\epsilon^{-1},$$

which, by (5.2), gives the fairly good prediction $\eta \leqslant 2\epsilon^{-1} \cdot u_* \doteq 2$.

B. Gaussian elimination with partial pivoting yields an L factor that satisfies $|L| \leqslant 1$ componentwise. This can be used (Higham, 2002, p. 143) to show that

$$\gamma(L,U) \leqslant \mathrm{cond}(L^{-1}) \leqslant 2^m - 1,$$

which proves that the growth factor remains bounded for a 'fixed' dimension m. However, the upper bound on $\mathrm{cond}(L^{-1})$ is attained for Wilkinson's famous matrix

$$A = \begin{pmatrix} 1 & & & 1 \\ -1 & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ -1 & \cdots & -1 & 1 \end{pmatrix}, \quad L = \begin{pmatrix} 1 & & & \\ -1 & \ddots & & \\ \vdots & \ddots & \ddots & \\ -1 & \cdots & -1 & 1 \end{pmatrix}.$$

[6] We write $a \doteq b$, if $a - b \approx u$.

Numerical experiments quickly exhibit very large backward errors:

```
>> m = 53; A = eye(m)-tril(ones(m),-1); A(:,m) = 1;
>> rand('seed',42); b = rand(m,1); x = A\b;
>> eta = norm(b-A*x,inf)/norm(A,inf)/norm(x,inf)

eta = 3.2342e-003
```

Our analysis yields a fairly good prediction

$$\eta = 3.23\cdots\times 10^{-3} \leqslant 2\gamma(L,U)\cdot u_* = 3.77\cdots\times 10^{-2}.$$

C. For symmetric positive-definite matrices, the solution of the linear system $Ax = b$ by 'Cholesky factorization' corresponds to the decomposition

$$f\colon A \overset{g}{\longmapsto} L \overset{h}{\longmapsto} x,$$

with $A = L\cdot L^{\mathrm{T}}$. A perturbation $\tilde{L} = L + \Delta L$ of the intermediate result by roundoff,

$$|\Delta L| \leqslant u\cdot|L|,$$

induces, as for (5.2), the backward error (with respect to the norm $\|\cdot\|_2$)

$$\eta \leqslant 2\frac{\||L|\cdot|L^{\mathrm{T}}|\|_2}{\|A\|_2}u_* = 2\gamma(L,L^{\mathrm{T}})u_*.$$

Since $\||L|\|_2 \leqslant \sqrt{m}\|L\|_2$ for any $m\times m$ matrix, we infer (Higham, 2002, p. 198)

$$\gamma(L,L^{\mathrm{T}}) \leqslant \frac{\||L|\|_2\||L^{\mathrm{T}}|\|_2}{\|A\|_2} \leqslant m\frac{\|L\|_2\|L^{\mathrm{T}}\|_2}{\|LL^{\mathrm{T}}\|_2} = m.$$

Hence, we have

$$\eta \leqslant 2mu_*,$$

which hints at the perfect normwise backward stability of the Cholesky method.

6. Model analysis of Gaussian elimination: the componentwise case

The matrix estimate (5.1) immediately yields an estimate of the 'componentwise' backward error

$$\begin{aligned}\omega = \max_j \frac{|\Delta A\cdot\tilde{x}|_j}{(|A|\cdot|\tilde{x}|)_j} &\leqslant \max_j \frac{(|\Delta A|\cdot|\tilde{x}|)_j}{(|A|\cdot|\tilde{x}|)_j} \leqslant 2\max_j \frac{(|L|\cdot|U|\cdot|\tilde{x}|)_j}{(|A|\cdot|\tilde{x}|)_j}u_* \\ &\leqslant 2\frac{\max_j(|L|\cdot|U|\cdot|\tilde{x}|)_j}{\min_j(|A|\cdot|\tilde{x}|)_j}u_* = 2\frac{\||L|\cdot|U|\cdot|\tilde{x}|\|_\infty}{\||A|\cdot|\tilde{x}|\|_\infty}\underbrace{\frac{\max_j(|A|\cdot|\tilde{x}|)_j}{\min_j(|A|\cdot|\tilde{x}|)_j}}_{=\sigma(A,\tilde{x})}u_*, \qquad (6.1)\end{aligned}$$

which by $U = L^{-1}A$, i.e. $|U| \leqslant |L^{-1}|\cdot|A|$, induces (Skeel, 1979, Theorem 4.4)

$$\omega \leqslant 2\mathrm{cond}(L^{-1})\sigma(A,\tilde{x})u_*. \qquad (6.2)$$

As our derivation shows, this is not necessarily the best possible concise bound, but it allows for the easy comparison with the normwise bound (with respect to $\|\cdot\|_\infty$)

$$\eta \leqslant 2\mathrm{cond}(L^{-1})u_*.$$

We see that the componentwise bound just differs by the additional factor $\sigma(A,\tilde{x}) \geqslant 1$. This factor measures the quality of the 'scaling' of the linear system with respect to $\tilde{x}$ and predicts an instability for badly scaled systems.

6.1 *Examples*

A. We return to the example B of Section 4.1. The growth factor and the scaling are given by

$$\mathrm{cond}(L^{-1}) = 3 + 4\epsilon, \quad \sigma(A,x) = 2 + 2\epsilon^{-1}.$$

Experimentally, for $\epsilon = 10^{-16}$, Gaussian elimination yields (partial pivoting is not used here because of $|L| \leqslant 1$)

$$\eta = 2.84\cdots \times 10^{-17} \leqslant 2\mathrm{cond}(L^{-1})u_* = 6.66\cdots \times 10^{-16}.$$

On the other hand, the componentwise backward error satisfies

$$\omega = 0.499\cdots \leqslant 2\mathrm{cond}(L^{-1})\sigma(A,x)u_* = 13.3\cdots .$$

Thus, the model analysis helps to understand the actual behaviour of the two error concepts. In particular, we see that scaling can be an issue for Gaussian elimination with partial pivoting if analysed componentwise.

B. There are matrices for which the upper bound (6.2) turns out to be too coarse. As an example, we consider totally positive matrix A such as the Hilbert matrix of the example A of Section 4.1 or matrices that appear in spline interpolation. These matrices factor with $L \geqslant 0$ and $U \geqslant 0$, these bounds being understood componentwise. Thus, it is best for us to stay with the following intermediate step in the chain of estimates (6.1):

$$\omega \leqslant 2\max_j \frac{(|L|\cdot|U|\cdot|\tilde{x}|)_j}{(|A|\cdot|\tilde{x}|)_j} u_*.$$

Here, we obviously have $|L|\cdot|U| = |A|$ and we can therefore directly infer the perfect stability estimate (de Boor & Pinkus, 1977)

$$\omega \leqslant 2u_*.$$

7. Model analysis of a single iterative refinement step

In this final section, we apply the model analysis to the understanding of the results (Skeel, 1980) on iterative refinement of Gaussian elimination. We recall that the iterative refinement of a calculated solution $\tilde{x}$ of a linear system $Ax = b$ consists of three steps: computing the residual $r_0 = b - A\tilde{x}$, solving $Aw = r_0$ for a calculated correction $\tilde{w}$ (reusing the LU decomposition of A) and updating $\tilde{y} = \tilde{x} + \tilde{w}$. If there were no roundoff errors in the refinement steps (i.e. $\tilde{w} = w$), we would obtain $\tilde{y} = x$, the exact solution.

In Sections 5 and 6, the model analysis of Gaussian elimination allowed roundoff errors just in the L and U factors of A, yielding some equivalent perturbation of that matrix. Because of the reuse of these factors in the iterative refinement step, we reasonably assume that both Gaussian elimination steps, i.e. those leading to $\tilde{x}$ and $\tilde{w}$, are affected by roundoff through a single perturbation $\tilde{A} = A + \Delta A$ satisfying the estimate (5.1). This way, the result $\tilde{y}$ of the iterative refinement is given by

$$\tilde{y} = \tilde{x} + \tilde{w}, \quad (A + \Delta A)\tilde{w} = r_0 = b - A\tilde{x} = \Delta A \cdot \tilde{x}.$$

The residual after this step is $r_1 = b - A\tilde{y} = r_0 - A\tilde{w} = \Delta A\tilde{w}$, and therefore

$$A\tilde{w} = r_0 - \Delta A\tilde{w} = \Delta A(\tilde{x} - \tilde{w}) = \Delta A(\tilde{y} - 2\tilde{w}), \quad \tilde{w} = A^{-1}\Delta A\tilde{y} - 2A^{-1}\Delta A\tilde{w}.$$

Hence, we have

$$|\Delta A\tilde{w}| \leqslant |\Delta A||A^{-1}||\Delta A||\tilde{y}| + 2|\Delta A||A^{-1}||\Delta A\tilde{w}|,$$

which by (5.1), i.e. $|\Delta A| \leqslant 2|L||U|u_* \leqslant 2|L||L^{-1}||A|u_*$, implies

$$\|\Delta A\tilde{w}\|_\infty \leqslant 4\,\mathrm{cond}^2(L^{-1})\,\mathrm{cond}(A^{-1}) \cdot \||A||\tilde{y}|\|_\infty u_*^2 + 4\,\mathrm{cond}(L^{-1})\,\mathrm{cond}(A^{-1})u_*\|\Delta A\tilde{w}\|_\infty.$$

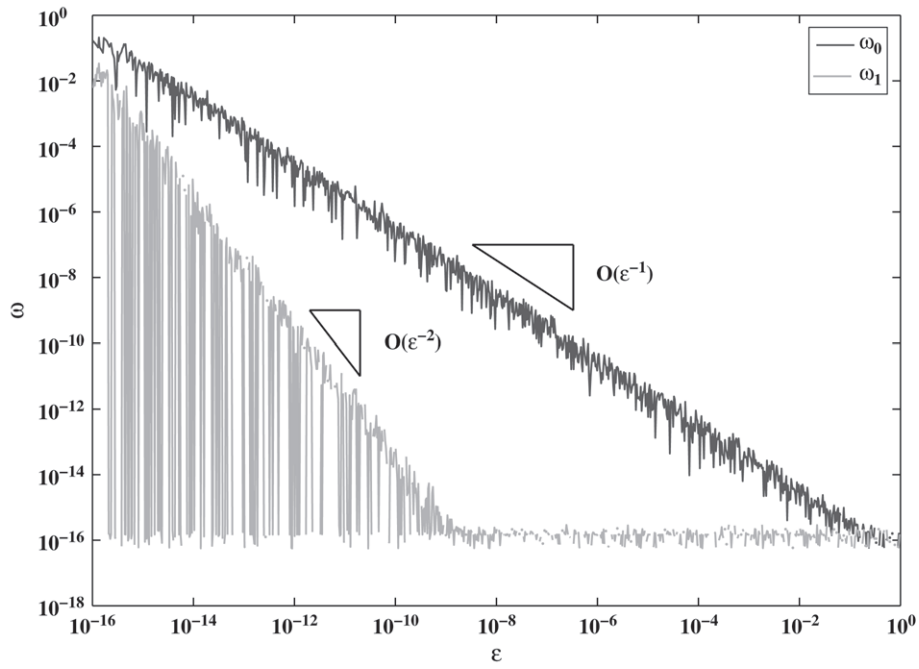

FIG. 1. Backward errors ω_0 and ω_1 versus ϵ.

If $4\,\mathrm{cond}(L^{-1})\,\mathrm{cond}(A^{-1})u_* < 1$, we can solve for $\|\Delta A w\|_\infty$ and get—as in the derivation of (6.1)—the following upper bound of the backward error of $\tilde{y}$:

$$\omega_1 = \max_j \frac{|r_1|_j}{(|A|\cdot|\tilde{y}|)_j} = \max_j \frac{|\Delta A\cdot\tilde{w}|_j}{(|A|\cdot|\tilde{y}|)_j} \leqslant \frac{\max_j |\Delta A\cdot\tilde{w}|_j}{\min_j (|A|\cdot|\tilde{y}|)_j}$$

$$= \frac{\|\Delta A\tilde{w}\|_\infty}{\||A||\tilde{y}|\|_\infty}\sigma(A,\tilde{y}) \leqslant \frac{4\,\mathrm{cond}^2(L^{-1})\mathrm{cond}(A^{-1})\sigma(A,\tilde{y})u_*}{1-4\,\mathrm{cond}(L^{-1})\mathrm{cond}(A^{-1})u_*}u_*. \qquad (7.1)$$

Because of $\mathrm{cond}(L^{-1}) \geqslant 1$ and $\sigma(A,\tilde{y}) \geqslant 1$, the premise is in particular satisfied if

$$8\,\mathrm{cond}^2(L^{-1})\mathrm{cond}(A^{-1})\sigma(A,\tilde{y})u_* \leqslant 1, \qquad (7.2)$$

for which we obtain from (7.1) the simple perfect bound $\omega_1 \leqslant u_*$. Except for a constant depending on the dimension m, this is 'exactly' the result (Higham, 2002, p. 239) of an elaborate analysis that takes all the details of roundoff error rigorously into account.[7]

In summary, our analysis predicts the following: As long as the linear system is not too badly conditioned (i.e. here, $\mathrm{cond}(A^{-1})$ is not too large) and not too badly scaled ($\sigma(A,\tilde{y})$ is not too large), and the Gaussian elimination is not too unstable ($\mathrm{cond}(L^{-1})$ is not too large), 'one step of iterative refinement implies componentwise backward stability'.

7.1 *Example*

We consider the example (Skeel, 1979, p. 500)

$$A = \begin{pmatrix} 3 & 2 & 1 \\ 2 & 2\epsilon & 2\epsilon \\ 1 & 2\epsilon & -\epsilon \end{pmatrix}, \quad b = \begin{pmatrix} 3+3\epsilon \\ 6\epsilon \\ 2\epsilon \end{pmatrix}, \quad x = \begin{pmatrix} \epsilon \\ 1 \\ 1 \end{pmatrix}$$

of a well conditioned (for this particular right-hand side b) but badly scaled linear system. Because of

$$\mathrm{cond}(A^{-1}) = \frac{6}{5}\epsilon^{-1} + \mathrm{O}(1), \quad \sigma(A,x) = \frac{3}{4}\epsilon^{-1} + \mathrm{O}(1), \quad \mathrm{cond}(L^{-1}) = \frac{8}{3} + \mathrm{O}(\epsilon),$$

Condition (7.2) reads as

$$1 \geqslant 8\,\mathrm{cond}^2(L^{-1})\,\mathrm{cond}(A^{-1})\sigma(A,\tilde{y})u_* = \frac{256}{5}\epsilon^{-2} + \mathrm{O}(\epsilon^{-1}),$$

i.e. one step of iterative refinement is predicted to imply stability as long as ϵ remains larger than about the square root of the unit roundoff

$$\epsilon \geqslant \frac{16}{5}\sqrt{5u} + \mathrm{O}(u^2) \approx 7.5 \times 10^{-8}.$$

In fact, the upper bound (6.2) predicts that the componentwise backward error ω_0 of $\tilde{x}$ behaves like $\omega_0 = \mathrm{O}(\epsilon^{-1}u)$, whereas the upper bound (7.1) predicts $\omega_1 = \mathrm{O}(\epsilon^{-2}u^2)$ for the first refinement step $\tilde{y}$. All these can be perfectly observed in an actual numerical experiment, see Fig. 1.

[7] The catch, of course, is that without doing the full analysis, we would not know if we had really determined the full bound. However, the point of this paper is a better understanding of the underlying mathematical structure. If, by neglecting many details, we come to predict the same bounds with much less effort, we seem to have put the focus on the right spot.

Acknowledgements

We are grateful to Nick Higham and Nick Trefethen for commenting on drafts of this paper.

REFERENCES

DE BOOR, C. & PINKUS, A. (1977) Backward error analysis for totally positive linear systems. *Numer. Math.*, **27**, 485–490.

HIGHAM, N. J. (2002) *Accuracy and Stability of Numerical Algorithms*, 2nd edn. Philadelphia, PA: SIAM.

OETTLI, W. & PRAGER, W. (1964) Compatibility of approximate solution of linear equations with given error bounds for coefficients and right-hand sides. *Numer. Math.*, **6**, 405–409.

RIGAL, J.-L. & GACHES, J. (1967) On the compatibility of a given solution with the data of a linear system. *J. Assoc. Comput. Machinery*, **14**, 543–548.

SKEEL, R. D. (1979) Scaling for numerical stability in Gaussian elimination. *J. Assoc. Comput. Machinery*, **26**, 494–526.

SKEEL, R. D. (1980) Iterative refinement implies numerical stability for Gaussian elimination. *Math. Comput.*, **35**, 817–832.

WILKINSON, J. H. (1963) *Rounding Errors in Algebraic Processes*. Englewood Cliffs, NJ: Prentice-Hall.