How to judge the error when you only know the residual.

It can be shown that the numerical solution $\hat{x}$ of the linear system $A\hat{x} = b$ obtained by Gauss elimination with partial pivoting satisfies exactly a ‘nearby’ linear system:

$$(A + E)\hat{x} = b,$$

with $\|E\| \approx \varepsilon \|A\|$. Since we are bound to change $A$ by a matrix $E$ of that size merely by converting its entries to floating-point numbers, this says that Gauss elimination with partial pivoting does as well as possible. But what does it say about error

$$e := x - \hat{x}$$

in our numerical approximation $\hat{x}$ to the (exact) solution $x := A^{-1}b$ of $A\hat{x} = b$?

We can compute its residual

$$r := b - A\hat{x}.$$

From the above, $r = E\hat{x}$, hence

$$\|r\| \approx \varepsilon \|A\|\|\hat{x}\|.$$  

But we can also compute $\|r\|$ explicitly. The only question is what this might tell us about $\|e\|$. Is the size of the relative residual

$$\|r\|/\|b\|$$

related to the size of the relative error

$$\|e\|/\|x\|?$$

By definition, the condition number, $\kappa(A)$, of a matrix $A$ is the greatest factor by which the relative error, $\|e\|/\|x\|$, can exceed the relative residual, $\|r\|/\|b\| = \|Ae\|/\|Ax\|$, i.e.,

$$\kappa(A) := \sup_{x,e} \frac{\|e\|/\|x\|}{\|Ae\|/\|Ax\|}.$$  

However, by interchanging here the roles of $x$ and $e$ and then taking reciprocals, this also says that

$$1/\kappa(A) = \inf_{e,x} \frac{\|e\|/\|x\|}{\|Ae\|/\|Ax\|}.$$  

Hence, altogether,

$$\frac{\|r\|}{\|b\|}/\kappa(A) \leq \frac{\|e\|}{\|x\|} \leq \kappa(A) \frac{\|r\|}{\|b\|}.$$  

1 ©2000 Carl de Boor
In other words, the larger the condition number, the less information about the size of the relative error is provided by the size of the relative residual.

If \( A \) is invertible, then we have the inequalities

\[
\|z\|/\|A^{-1}\| \leq \|Az\| \leq \|A\|\|z\|
\]

for any \( z \), and both inequalities are sharp, i.e., can be made equality by some choice of nonzero \( z \), hence then

\[
\kappa(A) = \|A\|\|A^{-1}\|.
\]

Apply this to the relative error in the approximate solution \( \hat{x} \) for \( A\hat{x} = b \) computed by Gauss elimination with pivoting. We conclude from (1) that its relative residual is \( \approx \varepsilon \|A\|\|\hat{x}\|/\|b\| \). Hence, with \( \|\hat{x}\| \approx \|x\| \) and \( \|x\|/\|b\| \leq \|A^{-1}\| \), we conclude that the relative error in \( \hat{x} \) is (approximately) bounded by \( \varepsilon \kappa(A) \), i.e.,

\[
\frac{\|x - \hat{x}\|}{\|x\|} \approx \varepsilon \kappa(A).
\]

This is beautifully illustrated in the textbook’s example on page 234, where the linear system

\[
A? := \begin{bmatrix}
.780 & .563 \\
.913 & .659
\end{bmatrix} = \begin{bmatrix}
.217 \\
.254
\end{bmatrix} =: [b]
\]

is considered. Its unique solution is \( x = (1, -1) \). The approximate solution \( \hat{x} = (.999, -1) \) has relative error \( \approx 10^{-3} \) and residual \( r = b - A\hat{x} = (.000780, .000913) \), hence relative residual \( \approx 10^{-3}/2 \), while the approximate solution \( \hat{x} = (.341, -0.087) \) has relative error \( \approx 1 \) and residual \( (.000001, 0) \), hence relative residual \( \approx 10^{-6} \). We conclude that the condition number for this matrix must be at least as big as \( 10^6 \). In fact,

\[
A^{-1} = 10^6 \begin{bmatrix}
.659 & -.563 \\
-.913 & .780
\end{bmatrix},
\]

showing that \( \kappa(A) \approx 10^6 \).

This also says that, for some different \( x \) (and corresponding \( b \)) and the same \( A \), we could have the other extreme situation, of a relative error that is only \( 10^{-6} \) of the relative residual.

We can visualize the condition number in the following way. Imagine the image

\[
\{Ax : \|x\| = 1\}
\]

of the ‘unit circle’ \( \{x : \|x\| = 1\} \) under the map \( A \). It will be some kind of ellipsoid, symmetric with respect to the origin. In particular, there will be a point \( x_{\text{max}} \) with \( \|x_{\text{max}}\| = 1 \) for which \( Ax_{\text{max}} \) will be as far away from the origin as possible. There will also be a point \( x_{\text{min}} \) with \( \|x_{\text{min}}\| = 1 \) for which \( Ax_{\text{min}} \) will be as close to the origin as possible. From the definition of matrix norms, it readily follows that \( \|Ax_{\text{max}}\| = \|A\| \), while \( \|Ax_{\text{min}}\| = 1/\|A^{-1}\| \). In particular,

\[
\kappa(A) = \|Ax_{\text{max}}\|/\|Ax_{\text{min}}\|,
\]
i.e., the condition number gives the ratio of the largest to the smallest diameter of the ellipsoid. The larger the condition number, the skinnier is the ellipsoid.

If now \( x = x_{\text{max}} \) while \( e = x_{\text{min}} \), then the relative error is 1 while the relative residual equals \( \|Ax_{\text{min}}\|/\|Ax_{\text{max}}\| \), and this is tiny to the extent that the ellipsoid is ‘skinny’. On the other hand, if \( x = x_{\text{min}} \) while \( e = x_{\text{max}} \), then the relative error is still 1 while now the relative residual equals \( \|Ax_{\text{max}}\|/\|Ax_{\text{min}}\| \), and this is large to the extent that the ellipsoid is ‘skinny’.

As another example, consider the matrix \( A = [x.^2 \ x \ \text{ones(size}(x)] \) for \( x = \text{linspace}(999,1001,101)' \). Using the max-norm for simplicity, we see that \( \|A*[1;0;0]\|_{\infty} \approx 10^6 \), with \( \|[1;0;0]\|_{\infty} = 1 \). To get some \( c \) of norm 1 with \( \|A\*c\|_{\infty} \) ‘small’, we make use of the fact that the three columns of \( A \) are almost linearly dependent, as can be seen by plotting its three columns as functions on the interval \([999 .. 1001]\). The straightforward plot \( \text{plot}(x,A) \) shows just two functions, both constant. The more discerning plot (shown below) is obtained by scaling the three columns to be of the same size: \( \text{plot}(x,[A(:,1)/A(1,1), \ A(:,2)/A(1,2), \ A(:,3)]); \) it shows two straight lines starting at the value 1 and, of course, the constant 1, hence suggests the following: with \( Ai=A(:,i)/A(1,i)-A(:,3), \ i=1:2 \), the vector \( A3 = A1 - (A1(101)/A2(101))\*A2 \) should be much smaller than either \( A1 \) or \( A2 \), yet it is of the form \( A\*c \) for some \( c \) of max-norm very close to 1. One computes that, in fact, \( \|A3\|_{\infty} \approx 10^{-6} \), hence this \( A \) has condition number at least \( 10^{12} \). Actually, \text{matlab} gives \( \text{cond}(A) = 3.2888e+012 \), hence our simple reasoning came very close to the extreme situations.
By contrast, the matrix $A = [(x-1000).^2 (x-1000) \text{ ones(size}(x)) ]$ made from appropriately shifted powers has $\text{cond}(A) = 3.7030$, a nice illustration of the fact that, when working with polynomials, one should at least work with appropriately shifted powers.

The worst condition number occurs when the columns of $A$ are linearly dependent, i.e., when there is a nonzero $x$ for which $Ax = 0$. For then, $\kappa(A) = \infty$ and the size of the relative residual says nothing about the size of the relative error. At the other extreme, the best $A$ to have is one for which $\kappa(A) = 1$. Now, whether this is so for a given $A$ depends strongly on just how we measure the vector norm. A major reason for using the Euclidean norm, $\| \|_2$, even though the corresponding matrix norm, $\|A\|_2$, is hard to compute, is the fact that there are many matrices for which $\kappa_2(A) = 1$. If $A$ is square, then these are the so-called unitary matrices, i.e., all square matrices $A$ for which

$$A' \cdot A = I,$$

hence also $A \cdot A' = I$.

Since $(A' \cdot A)(i,j) = A(:,i)' \cdot A(:,j)$, we see that the columns of a unitary matrix are orthogonal to each other and of Euclidean norm 1. Any matrix having this property, i.e., any matrix satisfying $A' \cdot A = I$, is therefore called orthonormal. Orthonormal matrices $A$ (whether square or not) have $\kappa_2(A) = 1$ and are therefore particularly desirable. They supply an orthonormal basis for their range. For an arbitrary matrix $A$, square or not, $\kappa_2(A) = \kappa_2(R1)$, with $A = Q1 \cdot R1$, and $Q1$ and $R1$ supplied by Matlab’s command $[Q1,R1]=\text{qr}(A,0)$. 

©2000 Carl de Boor