The residual again

The residual is our method of judging how good a potential solution $\tilde{\mathbf{x}}$ of a system $A\ \mathbf{x} = \mathbf{b}$ actually is. We compute

$$\mathbf{r} = \mathbf{b} - A\ \tilde{\mathbf{x}}$$

which gives us a measure of how good or bad $\tilde{\mathbf{x}}$ is as a potential solution.

One obvious complication of this idea is that a small residual does not necessarily mean that we are making a small mistake. A complicating factor is that in the original equation $A\ \mathbf{x} = \mathbf{b}$ the vector $\mathbf{b}$ on the right hand side establishes a natural scale for the problem. If $\|\mathbf{b}\|$ is large, we would expect solution vectors $\mathbf{x}$ to have similarly large norms. Likewise, if $\|\mathbf{b}\|$ is small, we would expect $\mathbf{x}$ to have a correspondingly smaller norm. The same reasoning applies to residuals. If $\|\mathbf{b}\|$ and $\|\mathbf{x}\|$ are both small, we would expect a typical residual $\mathbf{r}$ to have a small norm as well. In that case, simply having an $\mathbf{r}$ with small norm may not be sufficient. What matters more is the size of the norm $\|\mathbf{r}\|$ *relative to the natural scale induced on the problem by* $\|\mathbf{b}\|$. Likewise, what matters more to us than the size of the actual error $\|\mathbf{x} - \tilde{\mathbf{x}}\|$ is the *relative error*

$$\frac{\|\mathbf{x} - \tilde{\mathbf{x}}\|}{\|\mathbf{x}\|}$$

The next natural question to ask is what is the relationship between

$$\frac{\|\mathbf{r}\|}{\|\mathbf{b}\|}$$

which we can measure, and

$$\frac{\|\mathbf{x} - \tilde{\mathbf{x}}\|}{\|\mathbf{x}\|}$$

which gives us a true measure of the size of the error? The following theorem gives an answer.

Theorem Suppose that $\tilde{\mathbf{x}}$ is an approximate solution of the system $A\ \mathbf{x} = \mathbf{b}$, A is non-singular, and $\mathbf{r}$ is the residual vector associated with $\tilde{\mathbf{x}}$. Then, for any natural norm,

$$\|\mathbf{x} - \tilde{\mathbf{x}}\| \leq \|\mathbf{r}\|\ \|A^{-1}\|$$

and if $\mathbf{x} \neq \mathbf{0}$ and $\mathbf{b} \neq \mathbf{0}$,

$$\frac{\|\mathbf{x} - \tilde{\mathbf{x}}\|}{\|\mathbf{x}\|} \leq K(A)\ \frac{\|\mathbf{r}\|}{\|\mathbf{b}\|}$$

where

$$K(A) = \|A\|\ \|A^{-1}\|$$

is the *condition number* associated with the matrix A.

Proof From the definition of $\mathbf{r}$ we have

$$\mathbf{r} = \mathbf{b} - A\,\tilde{\mathbf{x}} = A\,\mathbf{x} - A\,\tilde{\mathbf{x}}$$

$$\mathbf{x} - \tilde{\mathbf{x}} = A^{-1}\,\mathbf{r}$$

taking norms on both sides and using the definition of the matrix norm of A gives

$$\|\mathbf{x} - \tilde{\mathbf{x}}\| = \|A^{-1}\,\mathbf{r}\| \leq \|A^{-1}\|\,\|\mathbf{r}\| \tag{1}$$

We also have

$$\mathbf{b} = A\,\mathbf{x}$$

$$\|\mathbf{b}\| = \|A\,\mathbf{x}\| \leq \|A\|\,\|\mathbf{x}\|$$

$$\frac{1}{\|\mathbf{x}\|} \leq \frac{\|A\|}{\|\mathbf{b}\|} \tag{2}$$

Multiplying inequality (1) by (2) gives

$$\|\mathbf{x} - \tilde{\mathbf{x}}\|\frac{1}{\|\mathbf{x}\|} \leq \|A^{-1}\|\,\|\mathbf{r}\|\,\frac{\|A\|}{\|\mathbf{b}\|}$$

or

$$\frac{\|\mathbf{x} - \tilde{\mathbf{x}}\|}{\|\mathbf{x}\|} \leq \|A\|\,\|A^{-1}\|\,\frac{\|\mathbf{r}\|}{\|\mathbf{b}\|}$$

Note In cases where the condition number $K(A)$ is large, a small relative residual

$$\frac{\|\mathbf{r}\|}{\|\mathbf{b}\|}$$

can correspond to a large relative error

$$\frac{\|\mathbf{x} - \tilde{\mathbf{x}}\|}{\|\mathbf{x}\|}$$

thus making the residual less useful as a predictor of success.

Since the theorem is valid for any natural norm, in most applications it suffices to use the norm that makes it easiest to compute the condition number $K(A)$. The text points out that in the $\|\ \|_\infty$ norm the matrix norm $\|A\|_\infty$ is easy to estimate:

$$\|A\|_\infty = \max_{1 \leq i \leq n} \left(|a_{i,1}| + |a_{i,2}| + \cdots + |a_{i,n}|\right)$$

Further insight

We can gain some further insight into what is going on here by using another important idea from linear algebra. If the matrix A is a real-valued matrix with n distinct eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$ and associated eigenvectors $\mathbf{v}_1$, $\mathbf{v}_2$,

…, $\mathbf{v}_n$ then any vector $\mathbf{v}$ can be written as a combination of those eigenvectors:

$$\mathbf{v} = c_1\,\mathbf{v}_1 + c_2\,\mathbf{v}_2 + \cdots + c_n\,\mathbf{v}_n$$

Expressing $\mathbf{v}$ as a combination of eigenvectors makes it easy to see what effect A will have on $\mathbf{v}$:

$$A\,\mathbf{v} = A\left(c_1\,\mathbf{v}_1 + c_2\,\mathbf{v}_2 + \cdots + c_n\,\mathbf{v}_n\right)$$

$$= c_1\,A\,\mathbf{v}_1 + c_2\,A\,\mathbf{v}_2 + \cdots + c_n\,A\,\mathbf{v}_n$$

$$= c_1\,\lambda_1\,\mathbf{v}_1 + c_2\,\lambda_2\,\mathbf{v}_2 + \cdots + c_n\,\lambda_n\,\mathbf{v}_n$$

Note that the eigenvectors are effectively rescaling the contributions from the various vectors $\mathbf{v}_j$. If λ_j is large in absolute value for some j, the contribution due to the term $c_j\,\mathbf{v}_j$ grows in importance after multiplying by A. Likewise, if λ_j is small in absolute value for some j, the contribution due to the term $c_j\,\mathbf{v}_j$ shrinks in importance after multiplying by A.

Now consider what happens to the norm $\|\mathbf{x} - \tilde{\mathbf{x}}\|$ as we pass from $\|\mathbf{x} - \tilde{\mathbf{x}}\|$ to $\|A\,\mathbf{x} - A\,\tilde{\mathbf{x}}\| = \|\mathbf{r}\|$. Suppose $\mathbf{x} - \tilde{\mathbf{x}}$ just happens to be strongly aligned with some eigenvector of A.

$$\mathbf{x} - \tilde{\mathbf{x}} = c_j\,\mathbf{v}_j$$

$$\mathbf{r} = A\,\mathbf{x} - A\,\tilde{\mathbf{x}} = A\,\left(c_j\,\mathbf{v}_j\right) = c_j\,\lambda_j\,\mathbf{v}_j$$

Taking norms gives us

$$\|\mathbf{r}\| = \|c_j\,\lambda_j\,\mathbf{v}_j\| = |\lambda_j|\,\|c_j\,\mathbf{v}_j\| = |\lambda_j|\,\|\mathbf{x} - \tilde{\mathbf{x}}\|$$

In the first scenario, the eigenvalue λ_j is large in absolute value. In that case, a large $\|\mathbf{r}\|$ corresponds to a small $\|\mathbf{x} - \tilde{\mathbf{x}}\|$. In the second scenario, the eigenvalue λ_j is small in absolute value. In that case, a small $\|\mathbf{r}\|$ can corresponds to a large $\|\mathbf{x} - \tilde{\mathbf{x}}\|$.

The bottom line here is that having one eigenvalue of A be small relative to the other eigenvalues of A can lead to the bad scenario of a small residual matched with a large relative error. That is exactly what the condition number $K(A)$ captures. It turns out that the condition number $K(A)$ is the ratio of the largest eigenvalue of A to the smallest eigenvalue of A.

The above discussion also shows us that not all errors are equally bad. If we happen to have an error term $\mathbf{x} - \tilde{\mathbf{x}}$ which is aligned with an eigenvector $\mathbf{v}_j$ with a relatively large eigenvalue λ_j of A, then a small residual really does correspond to a small error. On the other hand if the error term $\mathbf{x} - \tilde{\mathbf{x}}$ is aligned with an eigenvector $\mathbf{v}_j$ with a relatively small eigenvalue λ_j of A, then a small residual can correspond to a large error.

The method of iterative refinement

Here is one final application of the residual. Suppose we have just solved the system

$$A\,\mathbf{x} = \mathbf{b}$$

by Gauss elimination. Suppose we kept a record of the multipliers we used and can easily construct the L and U matrices for A.

Gauss elimination is subject to round-off errors, so we should not believe that the solution we computed is the exact solution of the system. Instead, we should treat it as an approximate solution $\tilde{\mathbf{x}}$ and apply our usual error analysis. Next, we compute the residual:

$$\mathbf{r} = \mathbf{b} - A\,\tilde{\mathbf{x}}$$

The *method of iterative refinement* takes this as a starting point and attempts to improve on the solution $\tilde{\mathbf{x}}$ via the following steps.

1. Use the L and U matrices to find an approximate solution $\tilde{\mathbf{y}}$ to the system

$$A\,\mathbf{y} = \mathbf{r}$$

2. Construct the vector

$$\tilde{\mathbf{x}} + \tilde{\mathbf{y}}$$

3. Use the latter vector in place of the original $\tilde{\mathbf{x}}$ you computed.

Why is this an improvement? Look at what happens when you multiply $\tilde{\mathbf{x}} + \tilde{\mathbf{y}}$ by A:

$$A\,(\tilde{\mathbf{x}} + \tilde{y}) = A\,\tilde{\mathbf{x}} + A\,\tilde{\mathbf{y}} \approx A\,\tilde{\mathbf{x}} + \mathbf{r} = A\,\tilde{\mathbf{x}} + \mathbf{b} - A\,\tilde{\mathbf{x}} = \mathbf{b}$$

The middle relation is only an approximate inequality, because $\tilde{\mathbf{y}}$ is only an approximate solution to

$$A\,\mathbf{y} = \mathbf{r}$$