The Condition Number for a Matrix

James Keesling

1 Condition Numbers

In the section we outline the general idea of a *condition number*. The condition number is a means of estimating the accuracy of a result in a given calculation. The simplest way to convey the idea is to do an example. Suppose that numbers on the computer are given with a certain accuracy ϵ. So, a given number x is represented roughtly by $x + \epsilon$. If the number x is represented to machine accuracy on a computer, then the IEEE standards require that $\frac{|\epsilon|}{|x|} \leq 2^{-52}$.

Suppose that we are evaluating a function f at the point x. Let us assume that we evaluate the function f completely accurately with the only error being the representation of the number x. Let $y = f(x)$ be the exact value of f at the true value of x and let $y + \delta = f(x + \epsilon)$. Then we have

$$\frac{f(x+\epsilon) - f(x)}{\epsilon} \approx \frac{\delta}{\epsilon} \approx f'(x).$$

We can say that $|f'(x)|$ is a condition number for the absolute error in the computation of f since $|\delta| \approx |f'(x)| \cdot |\epsilon|$. We can also obtain a condition number for the relative error in the computation. From the above inequality we have the following.

$$\begin{aligned} |\delta| &\approx |f'(x)| \cdot |\epsilon| \\ \frac{|\delta|}{|f(x)|} &\approx \frac{|f'(x)|}{|f(x)|}|x| \cdot \frac{|\epsilon|}{|x|} \end{aligned}$$

So, the condition number for the absolute error in computing f at x is $|f'(x)|$ and the condition number for the relative error in computing f at x is $\frac{|f'(x)|}{|f(x)|} \cdot |x|$.

2 Vector and Matrix Norms

To apply this to matrix computations, we will need a measure of the error. This is done by means of *vector norms* and *matrix norms* which we now describe.

Let V be an $n-$dimensional vector space. Let $||x||$ be a function from V to the non-negative real numbers, $||\cdot|| : V \to [0, \infty)$. Then $||x||$ is a norm for $x \in V$ if it satisfies the following axioms.

(1) $||x|| = 0$ if and only if $x = 0 \in V$.
(2) $||x + y|| \leq ||x|| + ||y||$ for all $x, y \in V$.
(3) $||\lambda \cdot x|| = |\lambda| \cdot ||x||$ for all $\lambda \in \mathbb{R}$ and $x \in V$.

The norm on an $n-$dimensional vector space is unique in the sense that if there are two norms $||x||_1$ and $||x||_2$, then there are positive real numbers r and s such that $r \leq \frac{||x||_2}{||x||_1} \leq s$ for all $x \neq 0$.

The norm with which you may be most familiar is called the *Euclidean norm.* Let $x = (x_1, x_2, \ldots, x_n)$. Then the Euclidean norm is given by

$$||x|| = \sqrt{\left(\sum_{i=1}^{n} x_i^2\right)}.$$

Another norm is the *sum norm* given by

$$||x|| = \sum_{i=1}^{n} |x_i|.$$

A third norm is the *maximum norm* given by

$$||x|| = \max_{1 \leq i \leq n} |x_i|.$$

There are other norms, but these three are the most important ones for an introductory course in numerical analysis. Although the Euclidean norm is the most familiar, the other two are easier to work with. The equivalence of norms on a finite-dimensional vector space justifies the use of norms that may perhaps be easier to work with.

Let $A : \mathbb{R}^n \to \mathbb{R}^n$ be a linear function. Suppose that $\mathbb{R}^n$ has the standard basis $\{e_1, e_2, \ldots, e_n\}$. Then A can be represented by an $n \times n$ matrix that we will also denote by A.

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2n} \\ a_{31} & a_{32} & a_{33} & \cdots & a_{3n} \\ \vdots & \vdots & & & \vdots \\ a_{n1} & a_{n2} & a_{n3} & \cdots & a_{nn} \end{bmatrix}$$

There is a *matrix norm* or *operator norm* on A that is defined in terms of the norm on $\mathbb{R}^n$ that is in current use. It is defined as follows

$$|||A||| = \max_{||x||=1} \frac{||Ax||}{||x||} = \max_{||x|| \neq 0} \frac{||Ax||}{||x||}.$$

3 Computing the Matrix or Operator Norm

All of the vector norms given above are relatively easy to compute. The difficulty comes in computing the operator norm for a given vector norm. For the Euclidean norm, we need to determine the largest norm on the image of the unit sphere under the operator. This requires that we find the largest semi-axis of an ellipsoid. It is much simpler to compute the operator norm for the sum norm and the maximum norm. These turn out to be quite simple formulas that are easily calculated.

Here is the operator norm for the maximum norm.

$$|||A||| = \max_{1 \le i \le n} \left\{ \sum_{j=1}^{n} |a_{ij}| \right\}$$

Here is the operator norm for the sum norm.

$$|||A||| = \max_{1 \le j \le n} \left\{ \sum_{i=1}^{n} |a_{ij}| \right\}$$

Here is a TI-89 program to compute these norms. The variable A is an $n \times n$ matrix. The variable n is the dimension of the matrix. The output is the variable $temp1$.

```
:maxnorm(A)
:Prgm
:dim(A)[1] → n
:0 → temp1
:For i,1,n
:∑(abs(A[i,j]),j,1,n) → temp
:max(temp1, temp) → temp1
:EndFor
:Disp temp1
:EndPrgm
```

To compute the operator norm for the sum norm, use the program above with A^{T} as the first argument. That will produce the maximum sum of the columns.

4 Significance of the Condition Number

The *condition number* of an $n \times n$ matrix A is

$$\mathrm{cond}(A) = |||A||| \cdot |||A^{-1}|||.$$

This number tells us how accurate we can expect the vector x when solving a system of equations $A \cdot x = b$. We assume that there is an error in representing the vector b, call it ϵ and otherwise the solution is given to absolute accuracy. That is we solve $A \cdot x = b + \epsilon$ and get a solution $x + \delta$ where x is the solution of $Ax = b$.

How does the condition number help estimate the number δ? We note that

$$x + \delta = A^{-1}(b + \epsilon) = A^{-1}b + A^{-1}\epsilon.$$

Since $A^{-1}b = x$, this gives us the following equation for δ.

$$\delta = A^{-1} \cdot \epsilon$$
$$||\delta|| \leq |||A^{-1}||| \cdot ||\epsilon||$$

So, the condition number for the magnitude of the absolute error δ for such a calculation is just the operator norm , $|||A^{-1}|||$.

On the other hand, the relative error is given by $\frac{||\delta||}{||x||}$. For the relative error we simply divide the above inequality by the norm of x to get the following inequality.

$$\frac{||\delta||}{||x||} \leq \frac{|||A^{-1}||| \cdot ||\epsilon||}{||x||}$$

However, from the definition of the norm of A, $\frac{||A \cdot x||}{||x||} \leq |||A|||$ and $Ax = b$. So, $||b|| \leq |||A||| \cdot ||x||$. Thus, combining these inequalities we get the following.

$$\frac{||\delta||}{||x||} \leq \frac{|||A^{-1}||| \cdot ||\epsilon||}{||x||} \leq \frac{|||A||| \cdot |||A^{-1}||| \cdot ||\epsilon||}{|||A||| \cdot ||x||} = \text{cond}(A) \cdot \frac{||\epsilon||}{||b||}$$

So, in solving the equation $Ax = b$, the relative error in the solution divided by the relative error in the right-hand-side vector is given by the condition number of A. The following rule of thumb is a useful way to express the above estimate. It states that if $m = \log_{10}(\text{cond}(A))$, then m is the number of digits accuracy lost in solving the system of equations $Ax = b$. There is typically additional error due to the many calculations needed in solving the equations. The estimate for additional losses is given by $\log_{10}(n)$ if the matrix A is $n \times n$.

5 Computing the Condition Number

Here is a simple TI-89 program to compute the condition number of an $n \times n$ matrix A using the operator norm associated with the maximum norm. The output variable is *condnum*.

```
:cond(A)
```

```
:Prgm
:dim(A)[1] → n
:maxnorm(A)
:temp1 → temp2
:maxnorm(A ∧ -1)
:temp1*temp2 → condnum
:Disp condnum
:EndPrgm
```

6 An Example

An ill-conditioned matrix is one whose condition number is large. A famous example is the Hilbert matrix. It is given by the following.

$$H = \begin{bmatrix} 1 & \frac{1}{2} & \frac{1}{3} & \cdots & \frac{1}{n} \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{4} & \cdots & \frac{1}{n+1} \\ \vdots & \vdots & & & \\ \frac{1}{n} & \frac{1}{n+1} & \frac{1}{n+2} & \cdots & \frac{1}{2n-1} \end{bmatrix}$$

Here is a program to enter the Hilbert matrix into your calculator. It becomes the matrix *hilb*.

```
:hilbert(n)
:Prgm
:newMat(n,n) → hilb
:For i,1,n
:For j,1,n
:1/(i+j-1) → hilb[i,j]
:EndFor
:EndFor
:EndPrgm
```

Here is the inverse of the Hilbert matrix for $n = 5$. Note that your calculator will compute the inverse exactly. The entries will be integers for all values of n.

$$H_5^{-1} = \begin{bmatrix} 25 & -300 & 1050 & -1400 & 630 \\ -300 & 4800 & -18900 & 26880 & -12600 \\ 1050 & -18900 & 79380 & -117600 & 56700 \\ -1400 & 26880 & -117600 & 179200 & -88200 \\ 630 & -12600 & 56700 & -88200 & 44100 \end{bmatrix}$$

The condition number for H_{10} is 3.536×10^{13}. On the TI-89, you cannot depend on any digits being correct solving a matrix equation $H_{10} \cdot x = b$.

7 Other Condition Numbers

As stated in §1, in general the *condition number* is a multiplier that relates a given error to one that would show up in an answer in a calculation. The condition number depends on the particular problem being solved. We can change the problem in the previous section to ask, "What error can we expect in solving for the inverse matrix for the matrix A?" Of course, in this case the given error will be in the way the matrix A is represented. Let us suppose that this error is Δ so that A is represented in the computer as $A + \Delta$.

Let us assume that we solve the equation $(A + \Delta)X = I$ exactly to get the solution $X + \delta$ where $X = A^{-1}$ is the true inverse of A. This gives us the following.

$$\begin{aligned} (A + \Delta)(X + \delta) &= I \\ AX + \Delta \cdot X + A\delta + \Delta \cdot \delta &= I \\ I + \Delta \cdot X + A\delta + \Delta \cdot \delta &= I \\ \Delta \cdot X + A\delta + \Delta \cdot \delta &= 0 \\ \Delta \cdot X &= -A\delta - \Delta\delta \\ -\delta &= \Delta X \cdot (A + \Delta)^{-1} \\ -\delta &= \Delta X \cdot (X + \delta) \end{aligned}$$

Let us assume that $|||A^{-1}||| = |||X||| \approx |||X + \delta|||$. Then we get the following result.

$$|||\delta||| \leq |||A^{-1}|||^2 \cdot |||\Delta|||$$

So, in this case the condition number for the error in computing the inverse of the matrix is $|||A^{-1}|||^2$. The relative error for this error is just $\frac{|||\delta|||}{|||A^{-1}|||}$. We simply divide the above inequality by $|||A^{-1}|||$ to get the following.

$$\frac{||\delta||}{|||A^{-1}|||} \leq |||A^{-1}||| \cdot |||\Delta|||$$