AM205: More on the condition number

Many numerical operations that we consider can essentially be boiled down to

$$y = f(x), \tag{1}$$

where x is a collection of some input values, y is a collection of some output values, and f is a function encapsulating the details of the operation. For any system such as this, an important numerical feature of interest is to know how a small change in the input from x to $x + \Delta x$ will affect the output. Mathematically, the change Δy to y is defined by

$$y + \Delta y = f(x + \Delta x). \tag{2}$$

Ideally, one would like that a small change in the input Δx would create only a small change in the output Δy, so that the numerical procedure is not sensitive to the initial conditions, and small errors in input will not create large variations in output. This can be mathematically characterized via the *condition number*, defined as

$$\kappa = \frac{|\Delta y / y|}{|\Delta x / x|}, \tag{3}$$

where the input and output are normalized so that x and y are dimensionless. Equation 3 is a rather loose, general definition and needs to be further specified depending on the situation. If x and y are vectors, then the $|\cdot|$ operators must be interpreted as some type of norm. In addition, κ will depend on the specific choices of x and Δx. Usually, the maximum bound on κ over the range of permissible values is reported.

The condition number for function evaluation

Suppose that x and y in Eq. 1 are scalars, and f is a real, differentiable function. Then by making use of Eq. 1 and 2,

$$\frac{\Delta y}{y} = \frac{f(x + \Delta x) - f(x)}{f(x)} = \frac{f(x + \Delta x) - f(x)}{\Delta x} \frac{\Delta x}{f(x)}. \tag{4}$$

Hence, if Δx is small,

$$\frac{\Delta y}{y} \approx \frac{f'(x)\Delta x}{f(x)}. \tag{5}$$

An approximate value of the condition number is therefore

$$\kappa \approx \left| \frac{f'(x)x}{f(x)} \right|. \tag{6}$$

As expected, the condition number is higher in places where f varies rapidly and f' is large, so that small changes in x will result in large changes in y.

The condition number for matrix calculations

Suppose that we now consider the condition number for the matrix multiplication

$$Ax = b, \tag{7}$$

where A is an invertible matrix, x is an input vector, and b is the output vector. Hence $A(x + \Delta x) = b + \Delta b$ and by linearity $A\Delta x = \Delta b$, so the condition number is given by

$$\kappa = \frac{\|\Delta b\| / \|b\|}{\|\Delta x\| / \|x\|} = \frac{\|A\Delta x\|}{\|\Delta x\|} \frac{\|x\|}{\|Ax\|} \tag{8}$$

where $\|\cdot\|$ represents any vector norm, such as the Euclidean norm. To proceed, a matrix norm can be defined in terms of the vector norm as

$$\|A\| = \max_{v \neq 0} \frac{\|Av\|}{\|v\|} \tag{9}$$

representing the maximum ratio that the matrix can scale a vector's length by. Then

$$\kappa \leq \|A\| \frac{\|x\|}{\|Ax\|}. \tag{10}$$

By rewriting $x = A^{-1}b$, this becomes

$$\kappa \leq \|A\| \frac{\|A^{-1}b\|}{\|b\|} \leq \|A\| \, \|A^{-1}\|, \tag{11}$$

and hence the upper bound on the condition number is the product of the matrix norm and the inverse matrix norm.

Suppose now that we consider closely-related problem of solving a linear system

$$Cy = f, \tag{12}$$

where C is an invertible matrix, f is the input vector of source terms, and y is the output solution. This can be rewritten as Eq. 7 by putting $C = A^{-1}$, $f = x$, and $y = b$. By following the same derivation as above, the condition number satisfies

$$\kappa \leq \|A^{-1}\| \, \|(A^{-1})^{-1}\| = \|C\| \, \|C^{-1}\|. \tag{13}$$

Therefore both problems—matrix muplication and solving a linear system—lead to exactly the same form of bound on the condition number.

As described previously, the condition number is often reported as a maximum bound over a range of values. Hence, the expression in Eq. 11 is often defined to be the condition number of a matrix,

$$\kappa(A) = \|A\| \, \|A^{-1}\|. \tag{14}$$

This can be computed using the `numpy.linalg.cond` function in Python, or the `cond` function in MATLAB.

Example for 2×2 diagonal matrices

Now suppose that the vector norm is given by the Euclidean norm. Consider a 2×2 invertible diagonal matrix of the form

$$A = \begin{pmatrix} \alpha & 0 \\ 0 & \beta \end{pmatrix} \tag{15}$$

where $|\alpha| \geq |\beta|$. Starting from Eq. 9, and writing v in terms of polar coordinates as $v = [r\cos\theta, r\sin\theta]^{\mathsf{T}}$, the matrix norm is

$$\begin{aligned}
\|A\| &= \max_{v \neq 0} \frac{\|Av\|}{\|v\|} \\
&= \max_{r \neq 0, \theta \in [0, 2\pi)} \frac{\sqrt{\alpha^2 r^2 \cos^2\theta + \beta^2 r^2 \sin^2\theta}}{\sqrt{r^2\cos^2\theta + r^2\sin^2\theta}} \\
&= \max_{\theta \in [0, 2\pi)} \frac{\sqrt{\alpha^2 \cos^2\theta + \beta^2 \sin^2\theta}}{\sqrt{1}} \\
&= \max_{\theta \in [0, 2\pi)} \sqrt{\alpha^2 - (\alpha^2 - \beta^2)\sin^2\theta}.
\end{aligned} \tag{16}$$

Since $\alpha^2 - \beta^2 \geq 0$, it follows that the expression will be maximized when $\theta = 0$, and hence

$$\|A\| = |\alpha|. \tag{17}$$

The inverse of the matrix is

$$A^{-1} = \begin{pmatrix} \alpha^{-1} & 0 \\ 0 & \beta^{-1} \end{pmatrix} \tag{18}$$

and applying the same argument shows that $\|A^{-1}\| = |\beta^{-1}|$. Hence

$$\kappa(A) = |\alpha\beta^{-1}|. \tag{19}$$

Note that while α and β also coincide with the eigenvalues of A for this particular example it not always the case that the condition number can be given in terms of the eigenvalues.