Conditioning and Numerical Stability

1 Introduction

Given a function f and a point x at which we wish to evaluate f, it is natural to ask how sensitive the output $f(x)$ is, in a relative sense, to a small perturbation of the input x. In other words, we are interested in quantifying the ratio

$$\frac{\text{Relative output error}}{\text{Relative input error}} = \frac{\|f(x+\Delta x)-f(x)\|/\|f(x)\|}{\|\Delta x\|/\|x\|}$$

where Δx represents some small perturbation of the input x. This is at the core of the theory of *conditioning*, a concept that was introduced in 1948 by Alan Turing (Turing 1948) and further developed in the 1960s by John Rice (Rice 1966). It is one of two important concepts that will be introduced in this note. The second concept is that of *numerical stability*, which is a desirable property of a numerical algorithm. Roughly speaking, numerical stability has to do with the accuracy of the output of an algorithm in the presence of errors introduced during the execution of the algorithm (e.g., rounding errors due to finite-precision arithmetic). It is important to distinguish between a *problem* and an *algorithm* for solving a problem: the condition number is a problem-specific property whereas numerical stability is a property of an algorithm.

2 The condition number

The condition number of a function $f\colon \mathbb{R}^n \to \mathbb{R}^m$ at a point $x \in \mathbb{R}^n$ may be defined as (Trefethen and Bau 1997)

$$\text{cond}(f,x) = \lim_{\varepsilon\to 0^+} \sup_{\|\Delta x\|\le\varepsilon} \frac{\|f(x+\Delta x)-f(x)\|/\|f(x)\|}{\|\Delta x\|/\|x\|}. \tag{1}$$

This is also known as the *relative* condition number of f, and it may be viewed as the best upper bound on the ratio of the relative output error to the relative input error for an infinitesimal perturbation Δx of the input x. The problem of evaluating f at x is said to be *well-conditioned* if $\text{cond}(f,x)$ is small. Roughly speaking, this means that $f(x)$ is insensitive to small input perturbations at x in a relative sense: relatively small changes in the input lead to relatively small changes in the output. On the other hand, if the condition number is large at x, then $f(x)$ is generally very sensitive to small input perturbations (again, in a relative sense) in which case the problem of evaluating f at x is said to be *ill-conditioned*. Finally, the problem of evaluating f at x is said to be *ill-posed* if the relative condition number is infinite.

2.1 Univariate functions

To develop an intuition about condition numbers, we briefly restrict our attention to univariate functions $f\colon \mathbb{R} \to \mathbb{R}$ that are twice continuously differentiable. From a Taylor expansion of such a function, we have that there exists a scalar $\theta \in (0,1)$ such that

$$f(x+\Delta x) = f(x) + f'(x)\Delta x + \frac{f''(x+\theta\Delta x)}{2}|\Delta x|^2.$$

This implies that

$$f(x+\Delta x) = f(x) + f'(x)\Delta x + o(|\Delta x|),$$

and, if we define $\Delta y = f(x+\Delta x) - f(x)$,

$$\Delta y \approx f'(x)\Delta x. \tag{2}$$

The absolute value of the derivative of f at x determines how sensitive, in an absolute sence, the output is to a sufficiently small input pertubation. This motivates the following definition of the so-called *absolute condition number* of a differentiable function f at x:

$$\text{cond}^{\text{abs}}(f,x) = |f'(x)|. \tag{3}$$

It is often more useful to quantify the sensitivity of our function f to input perturbations in a relative sense. Assuming that $x \neq 0$ and $f(x) \neq 0$, we can divide both sides of (2) by $f(x)$ which yields

$$\frac{\Delta y}{y} \approx \frac{xf'(x)}{f(x)}\frac{\Delta x}{x}, \tag{4}$$

where we used the fact that $y = f(x)$. This expression motivates the notion of a *relative condition number* which may be defined as

$$\text{cond}^{\text{rel}}(f,x) = \frac{|xf'(x)|}{|f(x)|} \tag{5}$$

for all x such that $f(x) \neq 0$. This definition coincides with (1) when f is univariate and continuously differentiable. We will often write $\text{cond}(f,x)$ instead of $\text{cond}^{\text{rel}}(f,x)$ when referring to the relative condition number. For a sufficiently small perturbation Δx, we have that

$$\frac{|\Delta y|}{|y|} \approx \text{cond}(f,x)\frac{|\Delta x|}{|x|}. \tag{6}$$

Example 1

The function $f(x) = \log(x)$ has the absolute condition number

$$\text{cond}^{\text{abs}}(f,x) = \frac{1}{|x|}$$

and the relative condition number

$$\text{cond}(f,x) = \frac{|x(1/x)|}{|\log(x)|} = \frac{1}{|\log(x)|}.$$

This implies that the problem of evaluating f at x is ill-conditioned when x is close to 1, and it is ill-posed when $x = 1$.

Example 2

Suppose $f(x) = a + x$ where a is a given constant. The absolute condition number of f at x is $\text{cond}^{\text{abs}}(f,x) = 1$, and the relative condition number is

$$\text{cond}(f,x) = \frac{|x|}{|a+x|}.$$

Thus, the problem of evaluating $f(x)$ at x is ill-conditioned for $x \approx -a$ (and ill-posed for $x = -a$).

Example 3

Suppose $f(x) = g(h(x))$ where $g\colon \mathbb{R} \to \mathbb{R}$ and $h\colon \mathbb{R} \to \mathbb{R}$ are continuously differentiable functions. From the chain rule of differentiation we have that $f'(x) = g'(h(x))h'(x)$, and hence

$$\text{cond}(f,x) = \frac{|g'(h(x))h'(x)x|}{|g(h(x))|} = \frac{|h(x)g'(h(x))|}{|g(h(x))|}\frac{|xh'(x)|}{|h(x)|} = \text{cond}(g,h(x))\text{cond}(h,x).$$

We will see later that this identity generally only holds for univariate functions.

Example 4

Consider the function $f(x) = x^{-1}$ for $x \neq 0$. The absolute condition number of f at x is given by $\text{cond}^{\text{abs}}(f,x) = 1/|x^2|$, and the relative condition number is

$$\text{cond}(f,x) = \frac{|x|/|x^2|}{1/|x|} = 1.$$

Thus, the problem of evaluating f at $x \neq 0$ is always well-conditioned.

Example 5

Consider the function $f(x) = \sqrt{x}$ for $x > 0$. The absolute condition number of f at x is $\text{cond}^{\text{abs}}(f,x) = 1/|2\sqrt{x}|$, and the relative condition number of f at x is

$$\text{cond}(f,x) = \frac{|x|/|2\sqrt{x}|}{\sqrt{x}} = \frac{1}{2},$$

which shows that the problem of evaluating f at $x > 0$ is always well-conditioned.

2.2 Indeterminate forms and L'Hôpital's rule

The relative condition number (5) is a ratio of two expressions, and it may happen that this is an *indeterminate form* such as $0/0$ or $0 \times \infty$ for some values of x. Clearly, $\text{cond}^{\text{rel}}(f,a)$ is an indeterminate form if $af'(a) = f(a) = 0$, in which case we may consider the limit

$$\lim_{x\to a} \frac{xf'(x)}{f(x)}.$$

To this end, we will make use of a result known as *L'Hôpital's rule*. To state the rule, we let g and h be differentiable functions on an open interval I that either contains the point a or has a as one of its endpoints. Now, suppose that for some $a \in \mathbb{R} \cup \{-\infty, +\infty\}$, we have that

$$\lim_{x\to a} g(x) = \lim_{x\to a} h(x) = 0 \qquad \text{or} \qquad \lim_{x\to a} |g(x)| = \lim_{x\to a} |h(x)| = \infty,$$

and

$$h'(x) \neq 0,\ \forall x \in I \setminus \{a\}, \qquad \lim_{x\to a} \frac{g'(x)}{h'(x)} \text{ exists},$$

then

$$\lim_{x\to a} \frac{g(x)}{h(x)} = \lim_{x\to a} \frac{g'(x)}{h'(x)}. \tag{7}$$

This is L'Hôpital's rule, and it is easy to prove in the special case where g and h are continuously differentiable on I and $a \in I$, i.e.,

$$\begin{aligned}\lim_{x\to a} \frac{g(x)}{h(x)} &= \lim_{x\to a} \frac{g(x)-g(a)}{h(x)-h(a)} = \lim_{x\to a} \frac{\frac{g(x)-g(a)}{x-a}}{\frac{h(x)-h(a)}{x-a}} = \frac{\lim_{x\to a} \frac{g(x)-g(a)}{x-a}}{\lim_{x\to a} \frac{h(x)-h(a)}{x-a}} \\ &= \frac{g'(a)}{h'(a)} = \lim_{x\to a} \frac{g'(x)}{h'(x)}.\end{aligned}$$

A proof of the more general case can be found in numerous books on calculus.

We note that indeterminate forms such as $0 \times \infty$ can be transformed to $0/0$ or ∞/∞ and evaluated using L'Hôpital's rule. For example, if $\lim_{x\to a} g(x) = 0$ and $\lim_{x\to a} h(x) = \infty$,

$$\lim_{x\to a} g(x)h(x) = \lim_{x\to a} \frac{g(x)}{1/h(x)}.$$

Example 6

Consider the function $f(x) = \log(a+x)$, defined for $x > -a$ with $a \neq 0$. The absolute condition number is $\text{cond}^{\text{abs}}(f,x) = 1/|a+x|$ and the relative condition number is

$$\text{cond}(f,x) = \frac{|x|}{|(a+x)\log(a+x)|}.$$

It is easy to verify that $\text{cond}(f,x) \to 0$ as $x \to \infty$, and using L'Hôpital's rule, we find that

$$\lim_{x\to -a} (a+x)\log(a+x) = \lim_{x\to -a} \frac{\log(a+x)}{1/(a+x)} = \lim_{x\to -a} \frac{1/(a+x)}{-1/(a+x)^2} = \lim_{x\to a} -(a+x) = 0,$$

which implies that $\text{cond}(f,x) \to \infty$ as $x \to -a$. Thus, we can conclude that the problem of evaluating f is ill-conditioned when x is near $-a$, and well-conditioned when x is sufficiently large.

Now, note that the denominator is equal to zero when $x = 1-a$, and hence the problem of evaluating f at $x = 1-a$ is ill-posed if $a \neq 1$. In the special case where $a = 1$, the numerator and the denominator are both equal to zero at $x = 1-a$, and using L'Hôpital's rule, we can evaluate the limit

$$\lim_{x\to 0} \frac{x}{(1+x)\log(1+x)} = \left.\frac{1}{\log(1+x) + \frac{1+x}{1+x}}\right|_{x=0} = 1.$$

Thus, the problem is well-conditioned at $x = 0$ when $a = 1$, and it can be evaluated using the function `log1p(x)`, which evaluates $\log(1+x)$ without computing $1+x$.

Example 7

Consider the function $f(x) = \int_0^1 e^{tx}\,dt$, which can also be expressed as

$$f(x) = \begin{cases} 1, & x = 0, \\ \frac{e^x - 1}{x}, & x \neq 0. \end{cases}$$

Using the Leibniz integral rule, we can express the derivative of f as

$$f'(x) = \int_0^1 \frac{\partial}{\partial x} e^{tx}\,dt = \int_0^1 te^{tx}\,dt,$$

or equivalently,

$$f'(x) = \begin{cases} \frac{1}{2}, & x = 0, \\ \frac{xe^x - x^x + 1}{x^2}, & x \neq 0. \end{cases}$$

It follows that the relative condition number is given by

$$\text{cond}^{\text{rel}}(f,x) = \begin{cases} 0, & x = 0, \\ \frac{|xe^x - e^x + 1|}{|e^x - 1|}, & x \neq 0. \end{cases}$$

Note that $\text{cond}^{\text{rel}}(f,x)$ is a continuous function of x since

$$\lim_{x\to 0} \frac{xe^x - e^x + 1}{e^x - 1} = \lim_{x\to 0} \frac{xe^x}{e^x} = 0,$$

and hence the problem of evaluating x near zero is well-conditioned.

We end this example by noting that it is a bad idea to evaluate $f(x)$ by directly computing $(e^x - 1)/x$ in finite-precision when x is near zero. Indeed, $\text{fl}(\text{fl}(e^x) - 1)$ may suffer from catastrophic cancellation. However, this can be avoided by using the function `expm1(x)`, which directly evaluates $h(x) = e^x - 1$ without computing e^x as an intermediate result.

2.3 Condition number of vector-valued and multivariate functions

The definition of the condition number of a univariate, real-valued function can be generalized to vector-valued and multivariate functions in several ways. If a function $f\colon \mathbb{R}^n \to \mathbb{R}^m$ is twice continuously differentiable and

$$f(x) = \begin{bmatrix} f_1(x) \\ \vdots \\ f_m(x) \end{bmatrix}$$

where $f_i\colon \mathbb{R}^n \to \mathbb{R}$, then a first-order Taylor approximation yields

$$f(x + \Delta x) = f(x) + J_f(x)\Delta x + o(\|\Delta x\|),$$

where $J_f(x) \in \mathbb{R}^{m\times n}$ denotes the Jacobian matrix which is defined as

$$J_f(x) = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_2}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \frac{\partial f_m}{\partial x_2} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix}.$$

Now, given a norm on $\mathbb{R}^m$, and letting $\Delta y = f(x+\Delta x) - f(x)$, it follows that

$$\|\Delta y\| \approx \|J_f(x)\Delta x\| \tag{8}$$

when Δx is sufficiently small. Recall that a subordinate matrix norm on $\mathbb{R}^{m\times n}$ induced by a vector norm on $\mathbb{R}^n$ is submultiplicative, i.e., if

$$\|J_f(x)\| = \sup_{u\neq 0}\left\{\frac{\|J_f(x)u\|}{\|u\|} : u \in \mathbb{R}^n\right\},$$

then it holds that

$$\|J_f(x)\Delta x\| \leq \|J_f(x)\|\|\Delta x\|.$$

Submultiplicativity implies that for a sufficiently small perturbation Δx, we have that

$$\|\Delta y\| \leqslant \|J_f(x)\|\|\Delta x\|. \tag{9}$$

This motivates the following generalization of the absolute condition number:

$$\text{cond}^{\text{abs}}(f,x) = \|J_f(x)\|. \tag{10}$$

Note that this definition depends on the choice of norm, and in the special case where $m = n = 1$, we have that $J_f(x) = f'(x)$.

The relative condition number (5) can be generalized to vector-valued and multivariate functions in a similar manner. It follows from (9) that if the perturbation Δx is sufficiently small and $f(x) \neq 0$, then

$$\frac{\|\Delta y\|}{\|y\|} \lessapprox \frac{\|x\|\|J_f(x)\|}{\|f(x)\|}\frac{\|\Delta x\|}{\|x\|}, \tag{11}$$

which motivates the following definition of the relative condition number:

$$\text{cond}^{\text{rel}}(f,x) = \frac{\|x\|\|J_f(x)\|}{\|f(x)\|}. \tag{12}$$

Note that this is a special case of (1) where f is assumed to be continuously differentiable. Like the absolute condition number, the relative condition number depends on the choice of norms, and in the special case where $m = n = 1$, it reduces to (5) for the Euclidean norm (and, in fact, for several other norms as well). We will use the shorthand notation $\text{cond}_p(f,x)$ to refer to the relative condition number of f at x in the p-norm (typically with p equal to 1, 2, or ∞).

Example 8: composite function

In this example, we derive the following property that applies to composite functions of the form $f(x) = g(h(x))$ where $g\colon \mathbb{R}^k \to \mathbb{R}^m$ and $h\colon \mathbb{R}^n \to \mathbb{R}^k$ are continuously differentiable:

$$\text{cond}(f,x) \leq \text{cond}(g,h(x))\text{cond}(h,x). \tag{13}$$

Using the chain rule of differentiation, the Jacobian of f, $J_f(x)$, may be expressed as

$$J_f(x) = J_g(h(x))J_h(x),$$

and hence

$$\begin{aligned}\text{cond}(f,x) &= \frac{\|J_f(x)\|\|x\|}{\|f(x)\|} = \frac{\|J_g(h(x))J_h(x)\|\|x\|}{\|g(h(x))\|} = \frac{\|J_g(h(x))J_h(x)\|\|h(x)\|\|x\|}{\|g(h(x))\|\|h(x)\|} \\ &\leq \underbrace{\frac{\|J_g(h(x))\|\|h(x)\|}{\|g(h(x))\|}}_{\text{cond}(g,h(x))} \underbrace{\frac{\|J_h(x)\|\|x\|}{\|h(x)\|}}_{\text{cond}(h,x)},\end{aligned}$$

where the inequality follows from the identity $\|J_g(h(x))J_h(x)\| \leq \|J_g(h(x))\|\|J_h(x)\|$.

Example 9: product

Let $f\colon \mathbb{R}^2 \to \mathbb{R}$ be defined as $f(x) = x_1x_2$. Then $J_f(x) = \begin{bmatrix} x_1 & x_2 \end{bmatrix}$, and hence the 2-norm absolute condition number of f at x is

$$\text{cond}_2^{\text{abs}}(f,x) = \|J_f(x)\|_2 = \sqrt{x_1^2 + x_2^2},$$

and the relative condition number is

$$\text{cond}_2(f,x) = \frac{x_1^2 + x_2^2}{|x_1x_2|} = |x_2/x_1| + |x_1/x_2|.$$

Thus, the problem of evaluating f at x is ill-conditioned when $|x_2/x_1|$ or $|x_1/x_2|$ is large, and it is ill-posed when x_1 or x_2 is equal to zero.

Example 10: fraction

Let $f(x) = x_1/x_2$. Then $J_f(x) = \begin{bmatrix} 1/x_2 & -x_1/x_2^2 \end{bmatrix}$, and hence the 2-norm absolute condition number of f at x is

$$\text{cond}_2^{\text{abs}}(f,x) = \|J_f(x)\|_2 = |x_2|^{-1}\sqrt{1 + (x_1/x_2)^2},$$

and the relative condition number is

$$\text{cond}_2(f,x) = \frac{\sqrt{x_1^2 + x_2^2}\|J_f(x)\|_2}{|x_1/x_2|} = \frac{1 + (x_1/x_2)^2}{|x_1/x_2|} = |x_2/x_1| + |x_1/x_2|.$$

Thus, the problem of evaluating f at x is ill-conditioned when $|x_2/x_1|$ or $|x_1/x_2|$ is large, and it is ill-posed when x_1 or x_2 is equal to zero.

Example 11: summation

Let $f(x) = x_1 + x_2 + \cdots + x_n$ for which the Jacobian is

$$J_f(x) = \begin{bmatrix} 1 & 1 & \cdots & 1 \end{bmatrix}.$$

We now derive the absolute and relative condition numbers of f based on the 1-norm. First, recall that the vector 1-norm on $\mathbb{R}^n$ is defined as $\|x\|_1 = \sum_{k=1}^n |x_i|$. The matrix 1-norm of a matrix $A \in \mathbb{R}^{p\times n}$ may be defined as

$$\|A\|_1 = \sup_{x\neq 0}\left\{\frac{\|Ax\|_1}{\|x\|_1}\right\} = \max_{k=1,\ldots,n}\left(\sum_{i=1}^p |A_{ij}|\right),$$

i.e., it is the maximum of the absolute column sums. This implies that $\|J_f(x)\|_1 = 1$, and hence $\text{cond}_1^{\text{abs}}(f,x) = 1$. The relative condition number is

$$\text{cond}_1(f,x) = \frac{\|x\|_1}{|f(x)|} = \frac{\sum_{i=1}^n |x_i|}{|\sum_{i=1}^n x_i|}.$$

If we use the infinity norm instead of the 1-norm, we get the relative condition number

$$\text{cond}_\infty(f,x) = \frac{n\|x\|_\infty}{f(x)} = \frac{n\max(|x_1|,\ldots,|x_n|)}{|\sum_{i=1}^n x_i|}.$$

Example 12: matrix-vector product

Let $f(x) = Ax$ where $A \in \mathbb{R}^{m\times n}$. The Jacobian of f is given by $J_f(x) = A$, and hence the 2-norm absolute and relative condition numbers are

$$\text{cond}_2^{\text{abs}}(f,x) = \|A\|_2, \quad \text{cond}_2(f,x) = \frac{\|A\|_2\|x\|_2}{\|Ax\|_2}.$$

We conclude that the problem of evaluating f is ill-posed for any $x \neq 0$ in the nullspace of A, and it is ill-conditioned when x is in the range of A and $\|Ax\|_2$ is small compared to $\|A\|_2\|x\|_2$.

Example 13: matrix-vector product

Let $f\colon \mathbb{R}^{n\times n} \to \mathbb{R}^n$ be defined as $f(A) = Ab$ where $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^n$. The absolute condition number in the matrix 2-norm is given by $\text{cond}_2^{\text{abs}}(f,A) = \|b\|_2$, and hence the relative condition number is

$$\text{cond}_2(f,A) = \frac{\|A\|_2\|b\|_2}{\|Ab\|_2}.$$

Example 14: matrix inverse

Let $f\colon \mathbb{R}^{n\times n} \to \mathbb{R}^{n\times n}$ be defined as $f(A) = A^{-1}$, i.e., f maps a nonsingular matrix to its inverse. In the matrix 2-norm, the absolute condition number of f at A is given by

$$\text{cond}_2^{\text{abs}}(f,A) = \|A^{-1}\|_2^2,$$

and the relative condition number is

$$\text{cond}_2(f,A) = \|A\|_2\|A^{-1}\|_2.$$

In linear algebra, the right-hand side is often denoted $\kappa(A)$ and is referred to as the condition number of A (with respect to inversion).

Example 15: sample variance

Let $x \in \mathbb{R}^n$ be a given vector, and let $f\colon \mathbb{R}^n \to \mathbb{R}$ be defined as

$$f(x) = \frac{1}{n-1}\|Cx\|_2^2, \quad n \geq 2,$$

where $C = I - \frac{1}{n}\mathbf{1}\mathbf{1}^T$ and $\mathbf{1}$ is the vector of ones. The matrix C is a so-called *centering matrix*,

and it is straightforward to verify that it satisfies $C^2 = C$ and $C\mathbf{1} = 0$. The product Cx is the vector x with its mean subtracted, i.e.,

$$Cx = x - \frac{1}{n}(\mathbf{1}^T x)\mathbf{1} = \begin{bmatrix} x_1 - \mu(x) \\ x_2 - \mu(x) \\ \vdots \\ x_n - \mu(x) \end{bmatrix}, \qquad \mu(x) = \frac{1}{n}\mathbf{1}^T x,$$

and hence $f(x)$ is the *sample variance* of x. The Jacobian of f is given by $J_f(x) = \frac{2}{n-1}x^T C^T C = \frac{2}{n-1}x^T C$, and hence the 2-norm condition number of f at x can be expressed as

$$\text{cond}_2(f,x) = \frac{\|x\|_2 \|J_f(x)\|_2}{|f(x)|} = 2\frac{\|x\|_2 \|Cx\|_2}{\|Cx\|_2^2} = 2\frac{\|x\|_2}{\|Cx\|_2}.$$

Noting that $x = \mu(x)\mathbf{1} + Cx$ and using the fact that $C\mathbf{1} = 0$, we have that $\|x\|_2^2 = n\mu(x)^2 + \|Cx\|_2^2$, and hence

$$\text{cond}_2(f,x) = 2\left(\frac{\|x\|_2^2}{\|Cx\|_2^2}\right)^{1/2} = 2\left(1 + n\frac{\mu(x)^2}{\|Cx\|_2^2}\right)^{1/2} = 2\left(1 + \frac{n}{n-1}\frac{\mu(x)^2}{f(x)}\right)^{1/2}.$$

3 Numerical stability

Numerical stability concerns the behavior of an algorithm when small errors are introduced during the computations. Roughly speaking, an algorithm is said to be *stable* if, for every input, such errors have an insignificant impact on the result, whereas it is said to be *unstable* if the impact is significant for one or more inputs.

Consider a function $y = f(x)$, and suppose that a given algorithm for evaluating $f(x)$ returns $\hat{y}$ as an approximation to y. The relative error is then $\|\hat{y} - y\|/\|y\|$. This is also referred to as the (relative) *forward error* of $\hat{y}$. The (relative) *backward error* of $\hat{y}$ is defined as

$$\inf_{\Delta x}\left\{\frac{\|\Delta x\|}{\|x\|} : \hat{y} = f(x + \Delta x)\right\},$$

i.e., it is the smallest relative perturbation of the input such that $\hat{y}$ is the exact solution to a nearby problem. Note that for a given x and $\hat{y}$, there may or may not exist a Δx such that $\hat{y} = f(x + \Delta x)$.

An algorithm is said to be *forward stable* if it always produces a result that satisfies

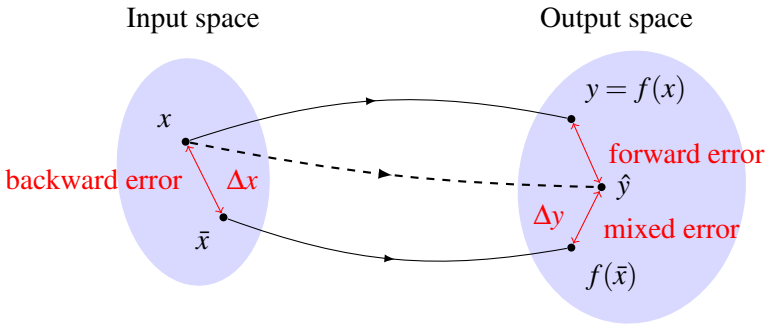
$$\|\hat{y} - y\| \leq c \cdot \text{cond}(f,x) \cdot \|y\|$$

for some small constant c. An algorithm is said to be *backward stable* if the backward error is always small (say, a small constant times the machine precision). Note that backward stability implies forward stability, but it does not guarantee that the forward error is small. Indeed, the forward error may be $\text{cond}(f,x)$ times larger than the backward error. In other words, the accuracy of the result depends on both the conditioning of the problem as well as the stability of the algorithm.

Finally, an algorithm is said to be *mixed forward-backward stable* if it always satisfies an error bound of the form

$$\hat{y} + \Delta y = f(x + \Delta x), \qquad \|\Delta y\| \leq \delta_1 \|y\|, \ \|\Delta x\| \leq \delta_2 \|x\|,$$

for some small constants δ_1 and δ_2. The following figure illustrates the different types of errors.



Example 16

Consider the function $f(x) = (1-\cos(x))/x^2$. The absolute condition number of f at x is

$$\text{cond}^{\text{abs}}(f,x) = \frac{|\sin(x)x - 2(1-\cos(x))|}{|x^3|},$$

and the relative condition number is

$$\text{cond}(f,x) = \frac{|\sin(x)x - 2(1-\cos(x))|}{|1-\cos(x)|}.$$

Both the numerator and the denominator are equal to 0 when $x = 0$, so we apply L'Hôpital's rule to find the limit

$$\lim_{x\to 0} \frac{\sin(x)x - 2(1-\cos(x))}{1-\cos(x)} = \lim_{x\to 0} \frac{\cos(x)x - \sin(x)}{\sin(x)} = \left.\frac{-\sin(x)x}{\cos(x)}\right|_{x=0} = 0.$$

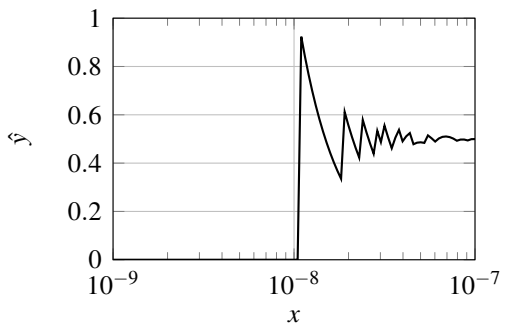
Thus, the problem of evaluating x at or near 0 is well-conditioned, but the naive implementation `(1-cos(x))/(x*x)` suffers from catastrophic cancellation when x is close to 0. To see this, first recall the cosine power series expansion

$$\cos(x) = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \cdots = \sum_{k=0}^{\infty} (-1)^k \frac{x^{2k}}{(2k)!}.$$

For a sufficiently small x, we have that

$$1 - \cos(x) \approx \frac{x^2}{2!},$$

and hence the limit of $f(x)$ as $x \to 0$ is $1/2$ (verify this using L'Hôpital's rule!). However, in double precision the naive implementation behaves quite differently, as shown in the following plot.



Example 17

Recall that $f(x) = \exp(x)$ can be expressed as

$$f(x) = \sum_{k=0}^{\infty} \frac{x^k}{k!} = 1 + x + \frac{x^2}{2} + \cdots$$

and note the recursive relationship between consecutive terms

$$\frac{x^k}{k!} = \frac{x^{k-1}}{(k-1)!}\frac{x}{k}, \quad k \geq 1.$$

This implies that $x^k/k! \to 0$ as $k \to \infty$.

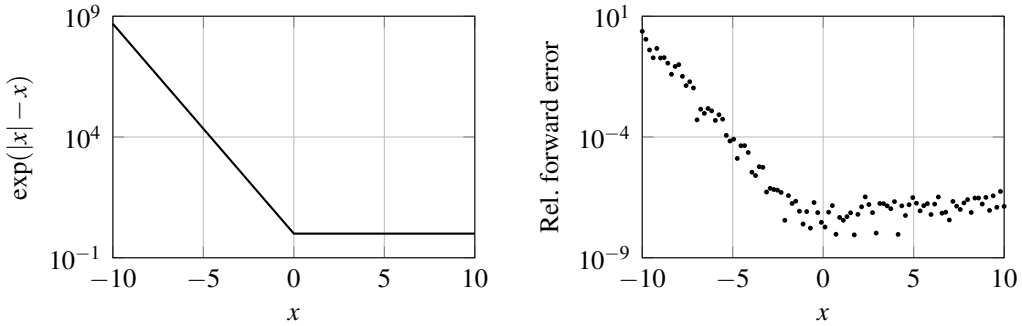
Now, suppose we want to evaluate $f(x)$ by summing the terms $x^k/k!$ for $k = 0, 1, 2, \ldots$ until $|x^k/k!|$ is so small that the finite-precision sum does not change. This can be implemented as follows:

Require: $x \in \mathbb{R}$
  $s \leftarrow 1, t \leftarrow 1, h \leftarrow 0, k \leftarrow 1$
  while $s \neq h$ **do**
    $h \leftarrow s$
    $t \leftarrow \text{fl}(t \,\text{fl}(x/k))$
    $s \leftarrow \text{fl}(s + t)$
    $k = k + 1$
  end while
  return s

The condition number of f is $\text{cond}(f, x) = |x|$ whereas the condition number for the summation of the series $1 + x + x^2/2 + \cdots$ may be expressed as

$$\frac{\sum_{k=0}^{\infty} |x^k|/k!}{\left|\sum_{k=0}^{\infty} x^k/k!\right|} = \frac{\exp(|x|)}{|\exp(x)|} = \exp(|x| - x).$$

The latter is equal to 1 when $x \geq 0$ (all terms are positive) but grows exponentially as x goes to $-\infty$. The following figure shows the condition number for the summation (left) and the relative forward error for our algorithm when the computations are carried out in single precision ($p = 24$) for different values of x in the range $[-10, 10]$ (right).



The algorithm is numerically unstable, but there is an easy fix: when x is negative, we can use the identity $x = |x| \cdot \text{sgn}(x)$ (where $\text{sgn}(x)$ denotes the sign function) to rewrite f as $f(x) = \exp(|x|)^{\text{sgn}(x)}$ in order to avoid using our algorithm with negative inputs.

References

Rice, John R. 1966. "A Theory of Condition." *SIAM Journal on Numerical Analysis* 3 (2): 287–310. https://doi.org/10.1137/0703023.

Trefethen, Lloyd N., and David Bau. 1997. *Numerical Linear Algebra*. Society for Industrial & Applied Mathematics (SIAM).

Turing, A. M. 1948. "Rounding-Off Errors in Matrix Processes." *The Quarterly Journal of Mechanics and Applied Mathematics* 1 (1): 287–308. https://doi.org/10.1093/qjmam/1.1.287.