§2 Machine Arithmetic

The arithmetic used on computers unfortunately does not respect the laws of ordinary arithmetic. Each elementary floating-point operation, in general, generates a small error that may then propagate through subsequent machine operations. As a rule, this error propagation is harmless, except in the case of subtraction, where cancellation effects may seriously compromise the accuracy of the results.

§2.1 A model of machine arithmetic

Any of the four basic arithmetic operations, when applied to two machine numbers, may produce a result no longer representable on the computer. We have therefore errors associated with arithmetic operations.

$\boxed{\text{Machine number}} \circ \boxed{\text{Machine number}} \longrightarrow$? machine number

$\Rightarrow$ errors associated with arithmetic operation

Barring the occurrence of overflow or underflow, we may assume a model of machine arithmetic as follows. Let x, $y \in \Re(t, s)$ and $fl(x \circ y)$ denotes the machine-produced result of the arithmetic operation $x \circ y$ then

$$fl(x \circ y) = x \circ y(1 + \epsilon),\ |\epsilon| \leq eps$$

This can be interpreted in a number of ways;

$$fl(x \times y) = [x(1 + \epsilon)] \times y = x \times [y(1 + \epsilon)] = (x\sqrt{1 + \epsilon}) \cdot (y\sqrt{1 + \epsilon})$$

Definition: Error propagation: A succession of arithmetic operations can well result in a significant error, owing to error propagation.

The rounding error are like germs which are widely spread. And only when weak spots are met, they will cause significant error.

§2.2 Error propagation in arithmetic operations: cancellation

We now study the extent to which the basic arithmetic operations propagate errors already present in their operatands. And assume that the arithmetic operations are exact.

(a) Multiplication:

$$\begin{aligned} x(1+\epsilon_x)\cdot y(1+\varepsilon_y) &= xy(1+\epsilon_x+\varepsilon_y+\varepsilon_x\varepsilon_y) \\ &\approx xy(1+\epsilon_x+\varepsilon_y) \end{aligned}$$

Thus, the relation error is given by $\varepsilon_{xy}=\varepsilon_x+\varepsilon_y$.
The relative errors in the data are being added to produce the error in the result. We consider this to be acceptable error propagation.

(b) Division ($y \neq 0$)

$$\begin{aligned} \frac{x(1+\epsilon_x)}{y(1+\varepsilon_y)} &= \frac{x}{y}(1+\varepsilon_x)(1-\varepsilon_y+\varepsilon_y^2+\cdots) \\ &\approx x/y(1+\epsilon_x-\varepsilon_y) \end{aligned}$$

$\therefore$ relation error $\varepsilon_{\frac{x}{y}} : \epsilon_x - \varepsilon_y$

(c) Addition and subtraction (assuming $x+y\neq 0$)

$$\begin{aligned} x(1+\epsilon_x)+y(1+\varepsilon_y) &= x+y+x\epsilon_x+y\varepsilon_y \\ &= (x+y)(1+\frac{x\epsilon_x+y\varepsilon_y}{x+y}) \end{aligned}$$

$\therefore \varepsilon_{x+y} = \frac{x}{x+y}\epsilon_x + \frac{y}{x+y}\varepsilon_y$

If $x\cdot y>0$ then $\varepsilon_{x+y}\leq |\epsilon_x|+|\epsilon_y|$

But what if $x+y\approx 0$ then $\varepsilon_{x+y}\uparrow$, and we call this as cancellation error. It is the only serious weakness – the Achilles heel of the numerical compuation, and it should be avoided whenever possible.

Example

x= $\boxed{1\ 0\ 1\ 1\ 0\ 0\ 1\ 0\ 1\ \mathbf{b}\ \mathbf{b}\ \mathrm{g}\ \mathrm{g}\ \mathrm{g}}$ $\boxed{\mathrm{e}}$

y= $\boxed{1\ 0\ 1\ 1\ 0\ 0\ 1\ 0\ 1\ b'\ b'\ \mathrm{g}\ \mathrm{g}\ \mathrm{g}}$ $\boxed{\mathrm{e}}$

x-y= $\boxed{0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ b"\ b"\ \mathrm{g}\ \mathrm{g}\ \mathrm{g}}$ $\boxed{\mathrm{e}}$

x-y= $\boxed{b"\ b"\ \mathrm{g}\ \mathrm{g}\ \mathrm{g}\ ?\ ?\ ?\ ?\ ?\ ?\ ?\ ?\ ?}$ $\boxed{\mathrm{e\text{-}9}}$

The cancellation phenomenon

Example: Quadratic equation:
Suppose $x^2 - 56x + 1 = 0$, find roots in 5-decimal arithmetic
$x_1 = 28 - \sqrt{783} = 28 - 27.982 = 0.018000$
$x_2 = 28 + \sqrt{783} = 28 + 27.982 = 55.982$
exact $x_1 = 0.0178628\cdots$ So, compute x_2 first then $x_1 = \frac{1}{x_2}$
$x_2 = 55.982137\cdots$

Example:
Compute $y = \cos(x+\delta) - \cos x = -2\sin(\frac{\delta}{2})\sin(x + \frac{\delta}{2})$ where $|\delta|$ is small

§3 The Condition of a Problem

A problem typically has an input and an output. The input consists of a set of data, say, the coefficients of some equation, and the output of another set of numbers uniquely determined by the input, say, all the roots of the equation in some prescribed order.

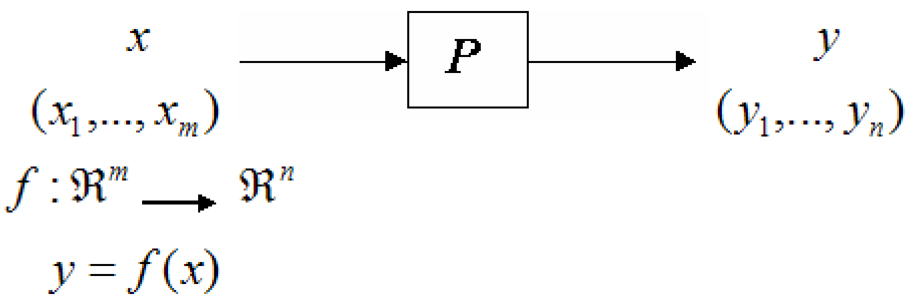


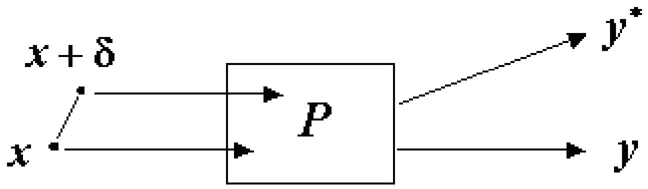
Figure 1: Black box representation of a problem

We may think of a problem as a map f, given by

$$f : \mathcal{R}^m \to \mathcal{R}^n,\ y = f(x).$$

What we are interested in is the sensitivity of the map f at some given point x to a small perturbation of x, that is how much bigger the perturbation in y is compared to the perturbation in x. In particular, we wish to measure the degree of sensitivity by a single number – the condition number of the map f at the point x. We emphasize that, as we pertub x, the function f is always assumed to be evaluated exactly, with infinitely precision. The condition of f, therefore, is an inherent property of the map f and does not depend on any algorithmic considerations concerning its implementation.

Definition: Condition measures the degree of sensitivity (assuming the evaluation of being precisely).



§3.1 Condition numbers

We start with the simplest case of a single function of one variable. Let $m = n = 1$, $x \neq 0$, $y \neq 0$.
$\Delta y = f(x + \Delta x) - f(x) \approx f'(x)\Delta x$
$\frac{\Delta y}{y} \approx \frac{f'(x)\Delta x}{f(x)} = \frac{xf'(x)}{f(x)} \cdot \frac{\Delta x}{x}$

Definition: Condition number of a function :
(cond f)(x) $:= \left|\frac{xf'(x)}{f(x)}\right|$
$x \neq 0,\ y = 0 \rightarrow \left|xf'(x)\right|$
$x = 0,\ y \neq 0 \rightarrow \left|\frac{f'(x)}{f(x)}\right|$
$x = 0,\ y = 0 \rightarrow \left|f'(x)\right|$
Proposition:
$\|A\|_\infty \stackrel{\text{def}}{\equiv} \max_{\substack{x\in\Re \\ x\neq 0}} \frac{\|Ax\|_\infty}{\|x\|_\infty}$, where $\|x\|_\infty = \sup_{1\leq\mu\leq m} |x_\mu|$

Then $\|A\|_\infty = \max_{1\leq\nu\leq n} \sum_{\mu=1}^{m} |a_{\nu\mu}| \leftarrow$ maximum row sum

Proposition: $\|y\|_1 = \sum_i |y_i|$, $\|A\|_1 \equiv \max_{\substack{x\in\Re \\ x\neq 0}} \frac{\|Ax\|_1}{\|x\|_1}$
Then $\|A\|_1 = \max_\mu \sum_\nu |a_{\nu\mu}| \leftarrow$ maximum column sum. Do this as an exercise.

Now, let's derive the global condition number for a vector-value function of several variables.

$$\Delta y_\nu = f_\nu(x + \Delta x) - f_\nu(x) \approx \sum_{\mu=1}^{m} \frac{\partial f_\nu}{\partial x_\mu}\Delta x_\mu$$

$$\Delta y_\nu \leq \sum_{\mu=1}^{m} \left|\frac{\partial f_\nu}{\partial x_\mu}\right| |\Delta x_\mu| \leq \max_\mu |\Delta x_\mu| \cdot \sum_{\mu=1}^{m} \left|\frac{\partial f_\nu}{\partial x_\mu}\right| \leq \max_\mu |\Delta x_\mu| \cdot \max_\nu \sum_{\mu=1}^{m} \left|\frac{\partial f_\nu}{\partial x_\mu}\right|$$

$$\therefore \|\Delta y\|_\infty \leq \|\Delta x\|_\infty \left\|\frac{\partial f}{\partial x}\right\|_\infty$$

$\therefore \frac{\|\Delta y\|_\infty}{\|y\|_\infty} \leq \frac{\|x\|_\infty \left\| \frac{\partial f}{\partial x} \right\|_\infty}{\|f\|_\infty} \cdot \frac{\|\Delta x\|_\infty}{\|x\|_\infty}$

So $(cond\ f)(x) := \frac{\|x\|_\infty \left\| \frac{\partial f}{\partial x} \right\|_\infty}{\|f\|_\infty}$

Example: System of linear algebraic equation:

$Ax = b$

$x = A^{-1}b = f(b)$

$(cond\ f)(b) = \frac{\|b\| \left\| A^{-1} \right\|}{\|A^{-1}b\|} = \frac{\|Ax\| \left\| A^{-1} \right\|}{\|x\|}$

$\max_{\substack{x \in \Re^n \\ x \neq 0}} \frac{\|Ax\| \left\| A^{-1} \right\|}{\|x\|} = \|A\| \, \|A^{-1}\|$

Example: Let's go back and look at the condition number

$$Df(x) = \begin{bmatrix} \frac{\partial f_1(x)}{\partial x_1} & \cdots & \frac{\partial f_1(x)}{\partial x_m} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_n(x)}{\partial x_1} & \cdots & \frac{\partial f_n(x)}{\partial x_m} \end{bmatrix}$$

$(cond\ f)(x) := \frac{\|x\|_\infty \|Df\|_\infty}{\|f\|_\infty}$, in ℓ_1 $\frac{\|x\|_1 \|Df\|_1}{\|f\|_1}$

This definition reduces precisely to earlier ones. But is much cruder than previous ones. This is because norms tend to destroy details. See the following definition.

Also, $(cond_{\nu\mu}\ f)(x) := \left| \frac{x_\mu \frac{\partial f_\nu}{\partial x_\mu}}{f_\nu(x)} \right| = \Gamma_{\nu\mu}$