Lesson 1

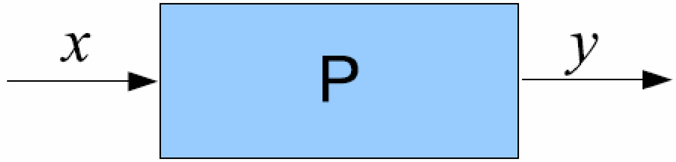
Floating Point System

Youndé – 6 August 2013

Proff. R. Bertelle – MR. Russo

A problem

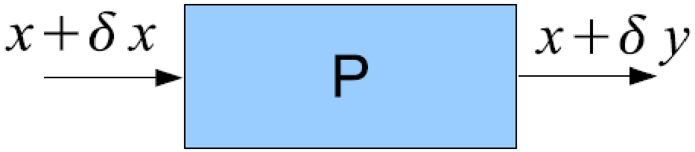
A problem $\mathcal{P}$ has inputs $x \in X$ and outputs $y \in Y$ where X and Y are some, normed, spaces for data x and solutions y, respectively. In an abstract manner, the problem $\mathcal{P}$ may be seen as a function $f : X \mapsto Y$.



(a) A problem P with inputs x and outputs y.

Conditioning of a problem

We say that the problem $\mathcal{P}$ with input x_0 is *well-conditioned* if all small, allowable, perturbations δx lead to small perturbations δy. Otherwise, if there is at least one small perturbation δx which leads to a large perturbation δy we say that the problem $\mathcal{P}$ with input x_0 is *ill-conditioned*.



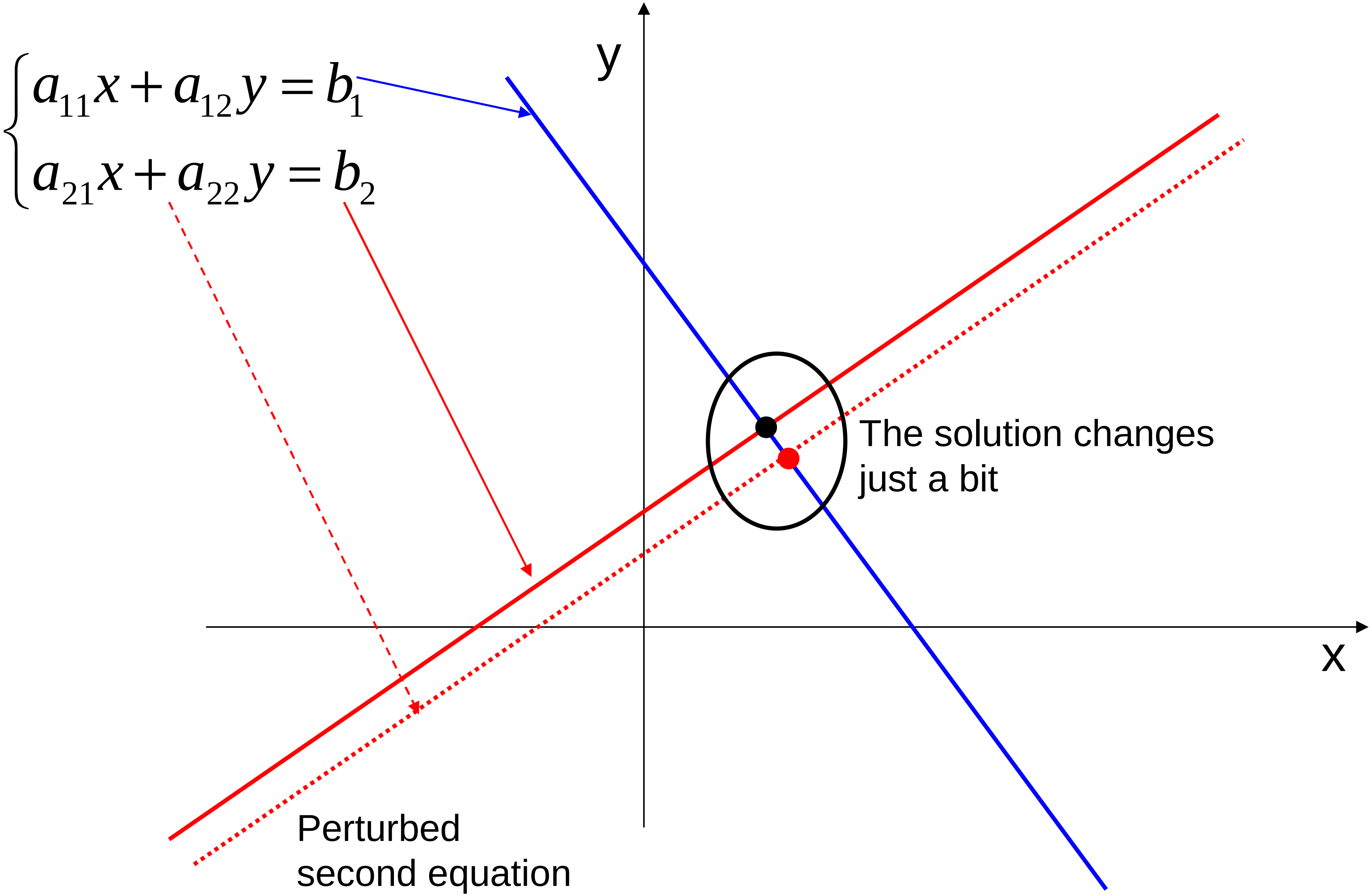
(b) A problem P with perturbed inputs $x + \delta x$ and the corresponding perturbed outputs $y + \delta y$.

One of the most useful, thought not the unique, measure for the conditioning of a problem $\mathcal{P}$ at x_0 is the *relative condition number* K. It is defined as

$$K = \sup_{\delta x} \frac{\left| \frac{\delta y}{y_0} \right|}{\left| \frac{\delta x}{x_0} \right|}$$

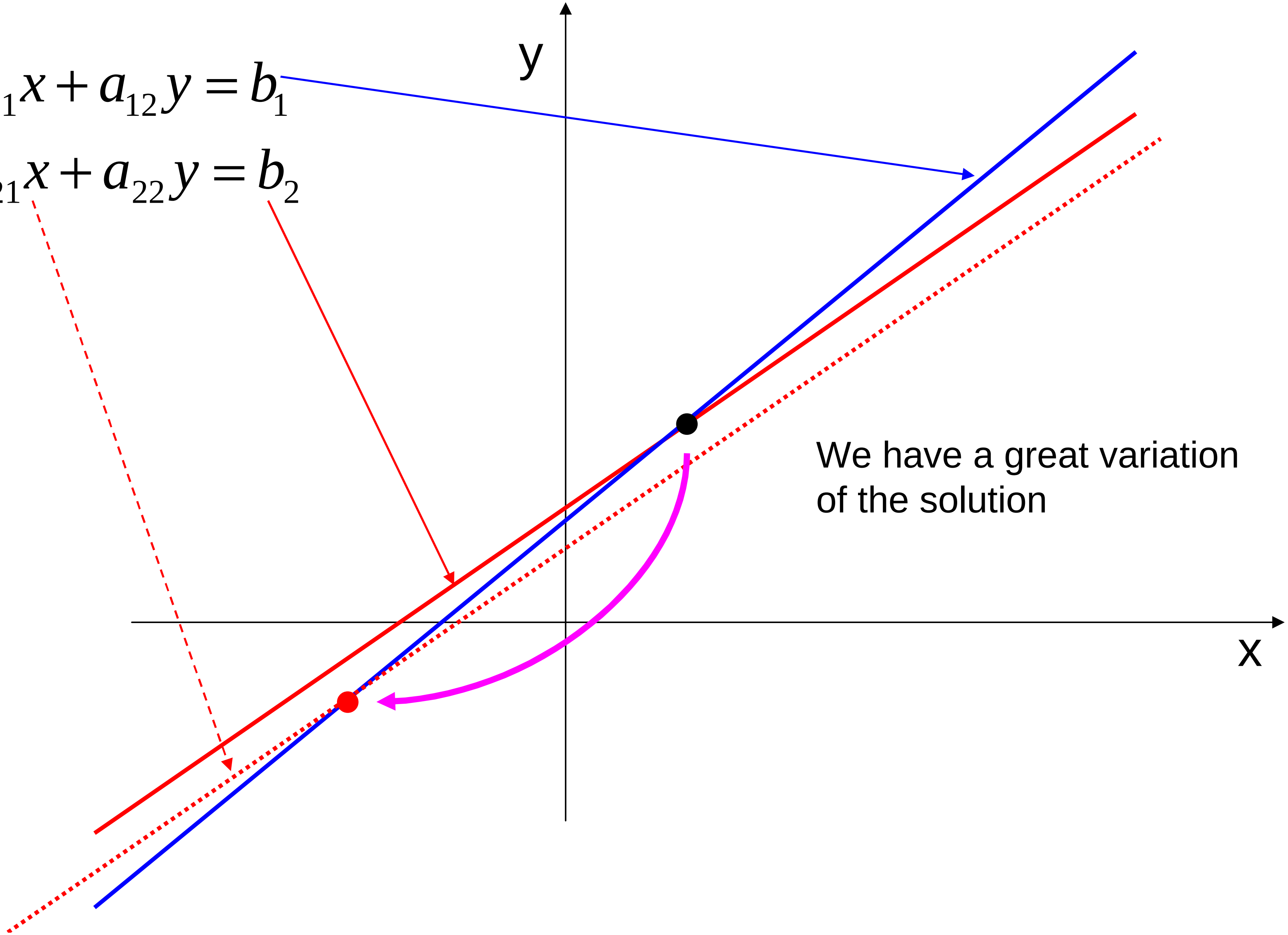
where the supremum is taken over all the allowable, small (infinitesimal from a mathematical point of view), perturbations δx. We say that the problem is well-conditioned if K is small, for example less than, about, 10^2; the problem is ill-conditioned if K is large, for example greater than 10^6.

Well-conditioned linear system



ill-conditioned linear system

$$\begin{cases} a_{11}x + a_{12}y = b_1 \\ a_{21}x + a_{22}y = b_2 \end{cases}$$



Conditioning of a function evaluation

Example 1.2 (Evaluation of a function) *Consider the computation of* $y_0 = f(x_0)$ *where* f *is a differentiable given function. Using Taylor expansion, we have*

$$f(x_0+\delta x) = f(x_0)+f'(x_0)\cdot\delta x+o(\ \delta x\ ) \qquad \Rightarrow \qquad \delta y = f(x_0+\delta x)-f(x_0) \approx f'(x_0)\cdot\delta x$$

and so, recalling that $y_0 = f(x_0)$, *we find*

$$\frac{\delta y}{y_0} = \frac{x_0 \cdot f'(x_0)}{f(x_0)} \cdot \frac{\delta x}{x_0} \qquad \Rightarrow \qquad K = \left| \frac{x_0 \cdot f'(x_0)}{f(x_0)} \right|$$

As an example, consider $f(x) = \sqrt{x+1} - \sqrt{x}$, $x \geq 0$. *Since the first derivative of* f *may be rewritten as*

$$f'(x) = \frac{-f(x)}{2\sqrt{x \cdot (x+1)}}$$

we obtain

$$K = \frac{|x_0|}{2\sqrt{x_0 \cdot (x_0+1)}}$$

and so the problem is well-conditioned for all $x_0 \geq 0$ *since* $K \leq 1/2$ *with* $K \approx 1/2$ *for* $x_0 \to +\infty$.

Conditioning of Eigenvalues Computation

Example 1.6 (Computation of the eigenvalues) *The computation of the eigenvalues of a non symmetric matrix is often an ill-conditioned problem. To see this, consider the matrices* A *and its perturbed version* $\hat{A}$ *defined as*

$$A = \begin{bmatrix} 101 & 110 \\ -90 & -98 \end{bmatrix} \qquad \hat{A} = \begin{bmatrix} 100 & 110 \\ -90 & -98 \end{bmatrix}$$

Note that the only difference among A *and* $\hat{A}$ *is that* $a_{11} = 101$ *and* $\hat{a}_{11} = 100$. *That is, a fairy small change, of the order of* 1%. *However, the eigenvalues of the two matrices are*

$$\begin{array}{lll} \lambda_1 = 1 & \lambda_2 = 2 & \textit{for matrix } A \\ \hat{\lambda}_1 \approx 1 + 10i & \hat{\lambda}_2 \approx 1 - 10i & \textit{for matrix } \hat{A} \end{array}$$

So, we have a large change in the eigenvalues a front of a small change in the matrix. Thus, according to our definition, the problem is ill-conditioned.

As a note, which we do not prove, the computation of the eigenvalues of a symmetric matrix is a well-conditioned problem. □

Floating Point Numbers #1

$$\mathbb{F}(\beta, t, L, U) = \{\ 0\ \} + \left\{ x \in \mathbb{R} \mid x = (-1)^s \cdot \beta^p \sum_{k=1}^{t} a_k\, \beta^{-k} \right\}$$

- β, the base, is an integer with $\beta \geq 2$. Common used bases are $\beta = 10$, $\beta = 2$ and $\beta = 16$.
- L and U are two integer numbers. Typically we have $L < 0 < U$. The scaling factor p is an integer satisfying $L \leq p \leq U$.
- t is a positive integer representing the number of figures a_k, $k = 1, \ldots, t$ of each floating point number. The unique representation of each floating point number requires $a_1 > 0$. Let us show what happens if this is not the case. Consider, as an example, the number $x = 1$ and $\mathbb{F}(10, 5, -6, 6)$. Then, the number $x = 1$ have different representations: 0.1×10^1, 0.01×10^2, 0.001×10^3 and many others.
- $s = 0$ for positive numbers and $s = -1$ for negative numbers.

Floating Point Numbers #2

$$\mathbb{F}(\beta, t, L, U) = \{\ 0\ \} + \left\{ x \in \mathbb{R} \mid x = (-1)^s \cdot \beta^p \sum_{k=1}^{t} a_k\, \beta^{-k} \right\}$$

Theorem 1.1 *The set of floating point numbers* $\mathbb{F}(\beta, t, L, U)$ *has the following properties.*

(a) $\mathbb{F} \subset \mathbb{R}$.

(b) if $x \in \mathbb{F}$ *then also* $-x \in \mathbb{F}$.

(c) $\mathbb{F}$ *has* $1 + 2 \cdot (\,\beta - 1\,) \cdot \beta^{t-1} \cdot (\,U - L + 1\,)$ *numbers.*

(d) The lower and the larger positive floating point numbers are, respectively, x_{min} *and* x_{max} *defined as*

$$x_{min} = \beta^{L-1}, \qquad x_{max} = \beta^U \cdot (\,1 - \beta^{-t}\,)$$

Floating Point Numbers #3

Example 1.7 *Let us explicitly write* $\mathbb{F}(10, 1, -1, 2)$*. It is* $\beta = 10$*,* $t = 1$*,* $L = -1$*,* $U = 2$*. Thus, for the positive floating point numbers, we have the 36 numbers shown in Table 1.1.*

$p = -1$	$p = 0$	$p = 1$	$p = 2$
$0.1 \cdot 10^{-1} = 0.01$	$0.1 \cdot 10^{0} = 0.1$	$0.1 \cdot 10^{1} = 1$	$0.1 \cdot 10^{2} = 10$
$0.2 \cdot 10^{-1} = 0.02$	$0.2 \cdot 10^{0} = 0.2$	$0.2 \cdot 10^{1} = 2$	$0.2 \cdot 10^{2} = 20$
$0.3 \cdot 10^{-1} = 0.03$	$0.3 \cdot 10^{0} = 0.3$	$0.3 \cdot 10^{1} = 3$	$0.3 \cdot 10^{2} = 30$
$0.4 \cdot 10^{-1} = 0.04$	$0.4 \cdot 10^{0} = 0.4$	$0.4 \cdot 10^{1} = 4$	$0.4 \cdot 10^{2} = 40$
$0.5 \cdot 10^{-1} = 0.05$	$0.5 \cdot 10^{0} = 0.5$	$0.5 \cdot 10^{1} = 5$	$0.5 \cdot 10^{2} = 50$
$0.6 \cdot 10^{-1} = 0.06$	$0.6 \cdot 10^{0} = 0.6$	$0.6 \cdot 10^{1} = 6$	$0.6 \cdot 10^{2} = 60$
$0.7 \cdot 10^{-1} = 0.07$	$0.7 \cdot 10^{0} = 0.7$	$0.7 \cdot 10^{1} = 7$	$0.7 \cdot 10^{2} = 70$
$0.8 \cdot 10^{-1} = 0.08$	$0.8 \cdot 10^{0} = 0.8$	$0.8 \cdot 10^{1} = 8$	$0.8 \cdot 10^{2} = 80$
$0.9 \cdot 10^{-1} = 0.09$	$0.9 \cdot 10^{0} = 0.9$	$0.9 \cdot 10^{1} = 9$	$0.9 \cdot 10^{2} = 90$

Considering also the negative ones and the zero we have

$$1 + 2 \cdot (\, U - L + 1\,) \cdot (\, \beta - 1\,) \cdot \beta^{t-1} = 1 + 2 \cdot [\, 2 - (-1) + 1\,] \cdot (\, 10 - 1\,) \cdot 10^{1-1} = 73$$

floating point numbers. Also, we have

$$x_{min} = 10^{L-1} = 10^{-1-1} = 0.01, \qquad x_{max} = 10^{U} \cdot (1 - 10^{-t}) = 10^{2} \cdot (1 - 10^{-1}) = 90$$

The difference between two consecutive numbers is not a constant. It is if they have the same value of p*.* $\square$

Converting real numbers into F #1

The positive real number x may be written, using the base β, as

$$x = \beta^p \sum_{k=1}^{+\infty} a_k \, \beta^{-k}$$

for some integer p and some non negative integers a_k with $a_1 \neq 0$. When this number has to be represented using a floating point number in the set $\mathbb{F}(\beta, t, L, U)$, one of the following cases may occur.

Converting real numbers into F #2

- If $p < L$ the number is less then the smallest representable floating point number. An *underflow* occurs.

- If $L \leq p \leq U$ the number can be represented on $\mathbb{F}$. There are, however, two cases

  - $a_k = 0$ for $k \geq t$. The number $x \in \mathbb{F}$ and so it can be exactly represented.
  - $a_k \neq 0$ for at least one $k > t$. The number $x \notin \mathbb{F}$. In this case, the better we can do is to represent the number x with the floating point number fl(x) (read: "the float of x") defined as

$$\mathrm{fl}(x) = \begin{cases} \beta^p \sum_{k=1}^{t} a_k \, \beta^{-k} & \text{if} \quad a_t \in \{0, \ldots, \frac{\beta}{2} - 1\} \\ \beta^p \sum_{k=1}^{t} a_k \, \beta^{-k} + \beta^{-t} & \text{if} \quad a_t \in \{\frac{\beta}{2}, \ldots, \beta - 1\} \end{cases}$$

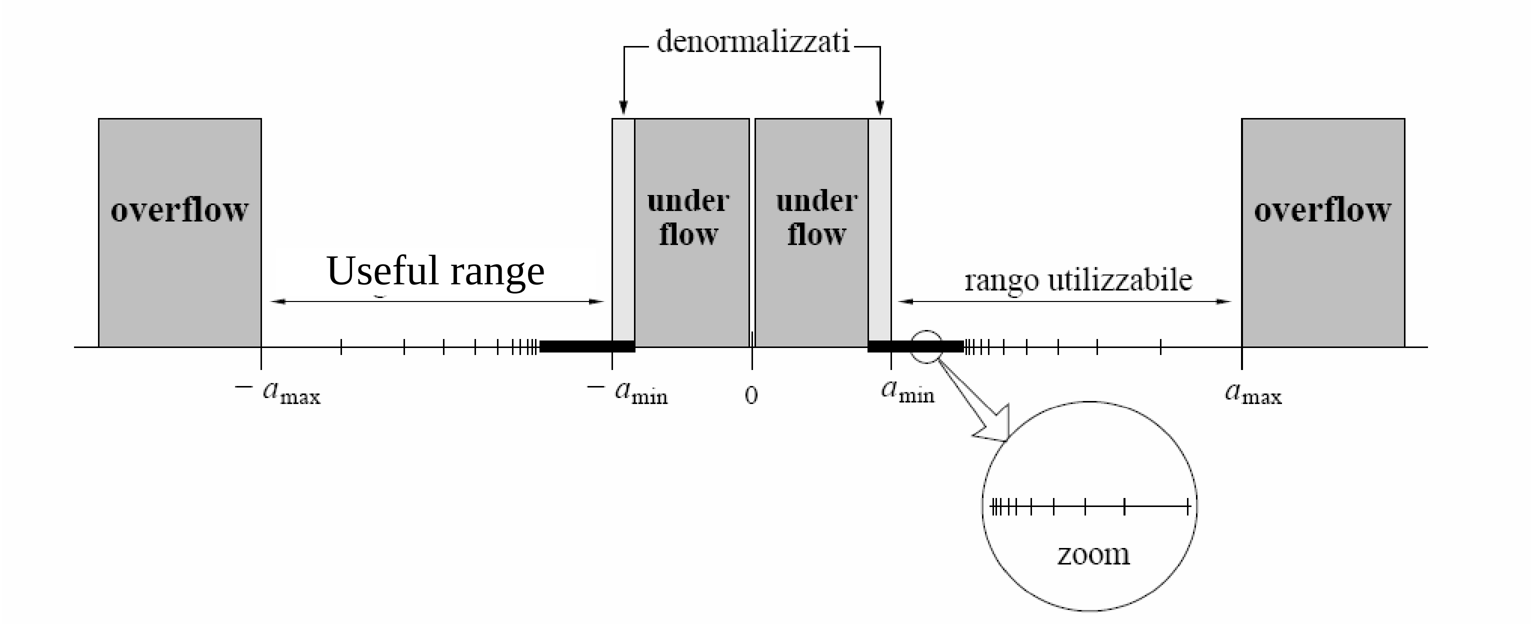
    The representation of $\mathrm{fl}(x)$ instead of x leads to an error called *rounding error*.

- $p > U$. The real number x is beyond the capacity of our floating point set $\mathbb{F}$. An *overflow* occurs and, usually, the computation stops with an error message.

Converting real numbers into F #3

Remark 1.2 (denormalized numbers) *Consider* $\mathbb{F}(\beta, t, L, U)$*. We have said that the first figure* a_1 *of each floating point number has to fulfill the condition* $a_1 > 0$ *in order to avoid multiple representations.*

However if, and only if, $e = L$ *it is usual to remove this condition allowing* a_1 *to be equal to zero. The real numbers obtained for* $e = L$, $a_1 = 0$ *and* $a_k \neq 0$ *for at least one* $k = 2, \ldots, t$*, are considered as new floating point numbers of* $\mathbb{F}$*. We call them* *__denormalized numbers__. The other numbers of* $\mathbb{F}$ *for which* $a_1 > 0$ *(regardless of* L*) are called* *__normalized numbers__.*



(Picture taken from book: M. Redivo Zaglia, "Calcolo Numerico, Metodi ed Algoritmi, 4 Edition")

Roundoff Error #1

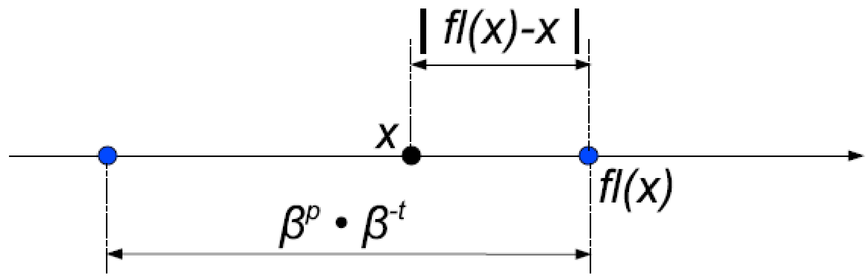
Theorem 1.2 *Let*

$$x = \beta^p \sum_{k=1}^{+\infty} a_k \, \beta^{-k}$$

be a positive, real number with $a_1 \neq 0$. *Then, assuming that there is non overflow, using the floating point system* $\mathbb{F}(\beta, t, L, U)$, *the following inequality holds*

$$\left| \frac{fl(x) - x}{x} \right| \leq \frac{\beta^{1-t}}{2} \tag{1.2}$$

Proof. Clearly, if $x \in \mathbb{F}$, we have $\mathrm{fl}(x) = x$ and thus $|\ \mathrm{fl}(x) - x\ | = 0$. So, the inequality is trivially fulfilled. Otherwise, the number x lies between two consecutive floating point numbers (blue circles in the next figure). The representative of x in $\mathbb{F}$ is the nearest to x of this two floating point numbers. As a consequence, it is $|\ \mathrm{fl}(x) - x\ | \leq \beta^{p-t}/2$.



Roundoff Error #2

Thus, recalling that $x > 0$ and so $|x| = x$, we have

$$\left| \frac{\mathrm{fl}(x) - x}{x} \right| = \frac{|\ \mathrm{fl}(x) - x\ |}{x} \overset{(1)}{\leq} \frac{|\ \mathrm{fl}(x) - x\ |}{\beta^p \cdot \beta^{-1}} \leq \frac{\frac{1}{2}\, \beta^p \cdot \beta^{-t}}{\beta^p \cdot \beta^{-1}} = \frac{\beta^{1-t}}{2}$$

where inequality (1) holds since (recall that $a_k \in \{0, 1, \cdots, \beta - 1\}$ and $a_1 > 0$)

$$\begin{aligned} x = \beta^p \sum_{k=1}^{+\infty} a_k\, \beta^{-k} &= \beta^p \left(a_1 \cdot \beta^{-1} + a_2 \cdot \beta^{-2} + a_3 \cdot \beta^{-3} + \cdots\right) \\ &\geq \beta^p \left(1 \cdot \beta^{-1} + 0 \cdot \beta^{-2} + 0 \cdot \beta^{-3} + \cdots\right) \\ &= \beta^p \cdot \beta^{-1} \end{aligned}$$

This ends the proof. $\square$

Machine precision #1

Definition 1.1 (machine precision) *Let* $\mathbb{F}(\beta, t, L, U)$ *be a floating point system. The number*

$$eps = \frac{\beta^{1-t}}{2} \tag{1.3}$$

is called the machine precision of the floating point system $\mathbb{F}$.

Machine precision #2

Note that 1 belongs to any floating point system since

$$1 = \beta^1 \cdot \beta^{-1} = \beta^1 \cdot \sum_{k=1}^{t} a_k \beta^{-k}$$

with $a_1 = 1$ and $a_k = 0$, $k = 2, \ldots, t$. The next floating point number is

$$x_+ = \beta^1 \cdot (\, 1 \cdot \beta^{-1} + 0 \cdot \beta^{-2} + \cdots + 0 \cdot \beta^{-t+1} + 1 \cdot \beta^{-t}\,) = \beta^1 \cdot (\, 1 \cdot \beta^{-1} + 1 \cdot \beta^{-t}\,)$$

which differs from 1 by $x_+ - 1 = \beta^{1-t} = 2\,\mathrm{eps}$. So, the real number $x = 1 + \mathrm{eps}$ lies exactly in the middle between 1 and x_+; thus, it is rounded to $\mathrm{fl}(1+\mathrm{eps}) = x_+$. Note also that each real number x satisfying $1 < x < 1 + \mathrm{eps}$ is rounded to the floating number 1.

From equation (1.2), for some $\bar{\epsilon}$ with $0 \leq \bar{\epsilon} \leq \mathrm{eps}$, we can write

$$\left| \frac{\mathrm{fl}(x) - x}{x} \right| = \bar{\epsilon} \quad \Leftrightarrow \quad \frac{\mathrm{fl}(x) - x}{x} = \pm\bar{\epsilon} \quad \Leftrightarrow \quad \mathrm{fl}(x) = x \pm \bar{\epsilon}\, x = x(1 \pm \bar{\epsilon})$$

Taking into account the sign, i.e. assuming $\epsilon \in [-\mathrm{eps}, \mathrm{eps}]$, $|\epsilon| = \bar{\epsilon}$, we have the following equation

$$\mathrm{fl}(x) = x\,(1 + \epsilon), \quad \epsilon \in [-\mathrm{eps}, \mathrm{eps}] \tag{1.4}$$

Floating Point Arithmetic #1

$$
\begin{array}{ll}
\text{(a)} & x \oplus y = \mathrm{fl}(x + y) \\
\text{(b)} & x \ominus y = \mathrm{fl}(x - y) \\
\text{(c)} & x \otimes y = \mathrm{fl}(x \times y) \\
\text{(d)} & x \oslash y = \mathrm{fl}(x - y)
\end{array}
\qquad x, y \in \mathbb{F}(\beta, t, L, U)
$$

So, each floating point operation require two steps: (i) execute the operation in $\mathbb{R}$; (ii) represent the obtained result in $\mathbb{F}$. As an example, consider $x \oplus y$.

(i) We first compute $x + y$ as an operation between the real numbers x and y.

(ii) We represent the result $x + y$ in $\mathbb{F}$ (considering, if the case, over and under flow).

Floating Point Arithmetic #2

Example 1.8 *Consider* $\mathbb{F}(10, -1, 2, 1)$ *and the three floating point numbers* $x = 0.1$, $y = 0.2$, $z = 0.7$. *Then, we have*

$$x \oplus y = fl(x + y) = fl(0.1 + 0.2) = fl(0.3) = 0.3$$

since $0.3 \in \mathbb{F}$. *Also, we have*

$$x \oslash z = fl(x/y) = fl(0.1/0.7) = fl(\ 0.14285714285714 \cdots\ ) = 0.1$$

Finally, $1 \oslash (x \otimes x)$ *gives an overflow; first, we compute*

$$x \otimes x = fl(x \times x) = fl(0.1 \times 0.1) = fl(0.01) = 0.01$$

next, we compute $1 \oslash 0.01 = fl(1/0.01) = fl(100)$; *since* 100 *is greater then the maximum representable floating point number in* $\mathbb{F}$, *an overflow is produced.* □

Floating Point Arithmetic #3

It is interesting to point out that most of the common properties of the operations

$x \oplus y = x$

if y is less then half of the distance between x and the next floating point number x_+.

Example 1.9 *Consider again* $\mathbb{F}(10, -1, 2, 1)$ *and the three floating point numbers* $x = 0.1$, $y = 2$, $z = 80$. *Using exact arithmetic, it is known that* $(x \times y) \times z = x \times (y \times z) = 16$. *Using floating point arithmetic, we have*

$$x \otimes y = fl(x \times y) = fl(0.1 \times 2) = fl(0.2) = 0.2$$

and

$$(x \otimes y) \otimes z = fl(0.2 \times 80) = fl(16) = 20$$

This is the best result we can have with our floating point system since $fl(x \times y \times z) = fl(16) = 20$. *On the other hand,* $x \otimes (y \otimes z)$ *returns an overflow since* $y \times z = 160$ *which is greater then the maximum representable number in* $\mathbb{F}$. *So, the executing order of the operations may be important.*

Floating Point Arithmetic #4

Example 1.10 (Smearing effect) *Consider the floating point system* $\mathbb{F}(10, 3, -2, 2)$ *and the three floating point numbers* $x = 0.123$, $y = 45.6$, $z = -45.5$. *The computation of* $x + y + z = 0.223$ *may be done in two ways.*

(i) We compute $w = x \oplus y$ *and then* $w \oplus z$. *We have*

$$w = x \oplus y \quad = \quad fl(0.123 + 45.6) = fl(45.723) = 45.7$$

and

$$w \oplus y \quad = \quad fl(45.7 - 45.5) = 0.200$$

(ii) We compute $u = y \oplus z$ *and then* $x \oplus u$. *We have*

$$u = y \oplus z = fl(45.6 - 45.5) = 0.100$$

and

$$x \oplus u = fl(0.123 + 0.100) = 0.223$$

So, in the first case the absolute value of the error is $0.10 = 10\%$ *whereas in the second case we have no error.*

Example 1.11 *Let* $f(x) = \sqrt{1+x} - \sqrt{x}$. *Consider the computation of* $f(49)$. *In exact arithmetic, we have* $f(49) = \sqrt{50} - \sqrt{49} = 0.07106781186548....$ *Using* $\mathbb{F}(10, -1, 2, 1)$ *and assuming that* $\sqrt{\xi}$ *is computed in a floating point system as* $fl(\sqrt{\xi})$, *we obtain*

$$fl(\sqrt{50}) = fl(7.07106781186548) = 7 \qquad \text{and} \qquad fl(\sqrt{49}) = fl(7) = 7.$$

Thus, using $\mathbb{F}(10, -1, 2, 1)$, *the obtained result is* $7 - 7 = 0$; *the absolute value of the relative error is* $|0.07106781186548... - 0|/0.07106781186548... = 1 = 100\%$.

Noting that

$$f(x) = \sqrt{1+x} - \sqrt{x} = \frac{(\sqrt{1+x} - \sqrt{x}) \cdot (\sqrt{1+x} + \sqrt{x})}{\sqrt{1+x} + \sqrt{x}} = \frac{1}{\sqrt{1+x} + \sqrt{x}}$$

we obtain

$$f(49) = fl\left(\frac{1}{\sqrt{50} + \sqrt{49}}\right) = fl\left(\frac{1}{7+7}\right) = fl(0.07142857142857...) = 0.07$$

which is the best possible result using this floating point system. The absolute value of the relative error is now

$$\frac{|0.07106781186548... - 0.07|}{0.07106781186548...} \approx 0.015 = 1.5\%$$

which is quite lower than the previous one.

Floating Point Arithmetic #7

Definition 1.2 (Stability of an algorithm) *An algorithm is stable if and only if small errors in the data and in the floating point operations does not grow up too much. Otherwise, the algorithm is instable.*

Floating Point Arithmetic #8

Example 1.13 *Consider the computation of the positive integrals*

$$I_n = \frac{1}{e}\int_0^1 x^n e^x dx, \quad n \in \mathbb{N} = \{0, 1, 2, \cdots\}$$

It is easy to see that $I_0 = 1 - e^{-1} = 0.6321205588285577\ldots$. *Moreover, integrating by parts, we get the recursive relation*

$$I_n = \frac{1}{e}\left\{ [\, x^n e^x \,]_0^1 - \int_0^1 n x^{n-1} e^x \right\} = 1 - nI_{n-1}.$$

Finally, it is easy to check that $\lim_{n\to+\infty} I_n = 0$ *since we have (recall that* $1 \le e^x \le e$, $x \in [0, 1]$*)*

$$0 \le \frac{1}{e}\int_0^1 x^n e^x dx \le \frac{1}{e}\cdot e \int_0^1 x^n dx = \frac{1}{n+1}$$

Now, consider the computation of I_n *for some given* $n > 1$ *with the following two algorithms:*

Floating Point Arithmetic #9

UNSTABLE

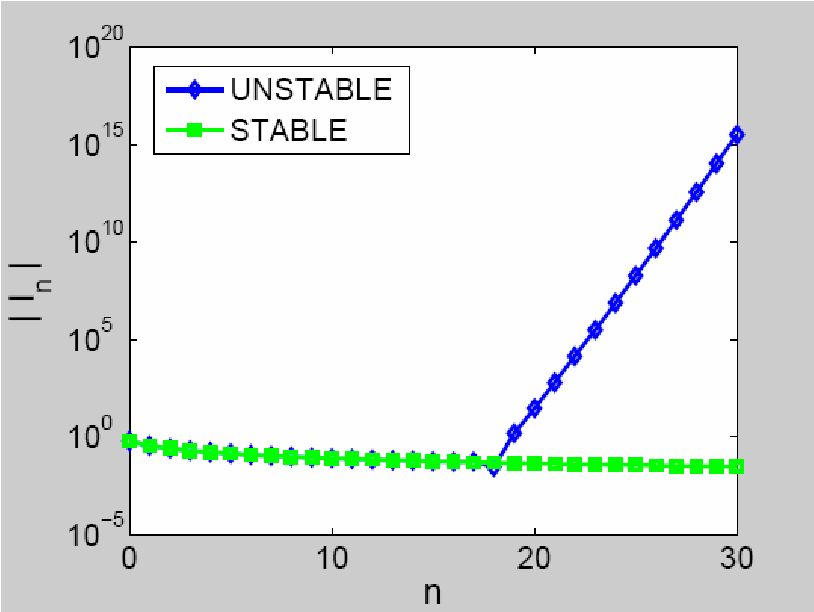
ALGORITHM 1

set $I_0 = 0.6321205588285577$
FOR k=1:n
$I_k = 1 - kI_{k-1}$
END

STABLE

ALGORITHM 2

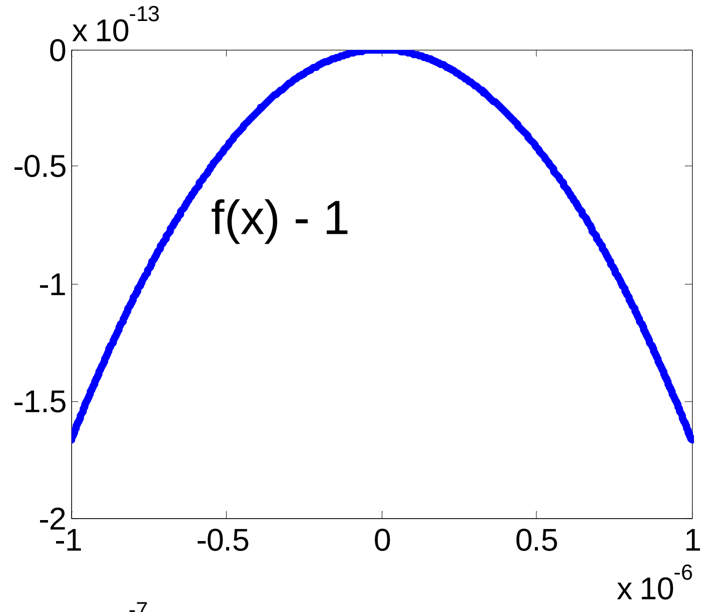
choose some N *with* $N \gg n$
set $I_N = 0$
FOR k=N:-1:n
$I_{k-1} = (1 - I_k)/k$
END



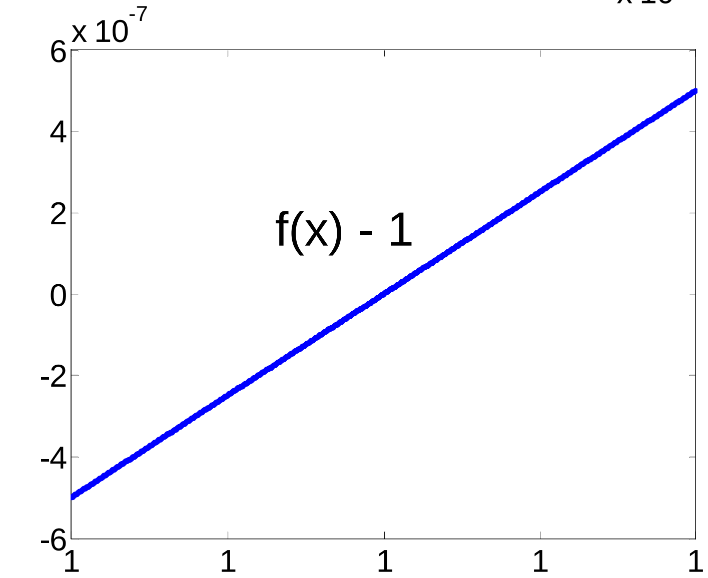
Floating Point Arithmetic #10

Some not so obvious (and not
so easy to prove) **stable formulas**.
(from "Introduction to Numerical Analysis"
Arnold Neumaier, Cambridge)

$$f(x) := \begin{cases} 1 & \text{if } x = 0 \\ \sin x / x & \text{if } x \neq 0 \end{cases}$$



$$f(x) := \begin{cases} 1 & \text{if } x = 1 \\ (x-1)/\ln x & \text{if } x \neq 1 \end{cases}$$



Floating Point Arithmetic #10

How to rewrite some unstable formulas in a stable way

unstable per large x

$$\sqrt{x+1} - \sqrt{x} = \frac{1}{\sqrt{x+1} + \sqrt{x}},$$

unstable for x near zero

$$1 - \cos x = \frac{\sin^2 x}{1 + \cos x} = 2 \sin^2 \frac{x}{2},$$