Calculatrices, nombres positifs, précisions relatives, pourcentages et distance logarithmique

Stéphane Junca(*)

1. Introduction

La représentation des nombres utilisée par la calculatrice modifie grandement la notion de précision utilisée dans nos classes de Mathématiques.

En effet, nos calculatrices utilisent des nombres flottants. C'est-à-dire que pour un nombre $x > 0$, on utilise son écriture à virgule flottante : $x = m10^p$, où m est la mantisse et p est l'exposant tels que : $1 \leq m < 10$ et $p \in \mathbf{Z}$. De plus, comme la capacité mémoire de la calculatrice est finie, la calculatrice ne dispose que d'une dizaine de décimales de la mantisse et l'exposant est souvent compris entre -99 et $+99$. On obtient ainsi un ensemble *fini* de nombres. Cet ensemble particulier et fini de nombres est l'ensemble des nombres flottants de la calculatrice. Cette représentation scientifique des nombres sera rappelée plus en détail dans la partie 3.

La calculatrice utilise parfois des nombres flottants à l'affichage pour les nombres trop grands ou trop petits. En revanche elle les utilise toujours en représentation interne et en calcul interne. Les puissances de 10 sont les nombres flottants les plus simples et ils se rencontrent dès le Collège. Nous allons voir que l'utilisation de ces nombres flottants modifie naturellement notre notion « euclidienne » de précision. En effet, un utilisateur d'une calculatrice demande que le résultat affiché soit le plus précis possible. Pour cela, il veut savoir combien de chiffres consécutifs à partir du premier chiffre affiché à gauche sont exacts. Il s'intéresse donc au *nombre de chiffres significatifs*, et donc, sans s'en apercevoir, comme on va l'expliciter plus loin, il ne s'intéresse finalement qu'à la *précision relative* du nombre affiché. On va quitter ainsi notre bonne distance euclidienne pour des distances relatives. Les distances relatives vérifient toutes une propriété géométrique fondamentale : elles sont toutes *invariantes par changement d'échelle*. Ceci est très important dans la pratique, car cela signifie que le résultat doit être invariant par changement d'unité. Mais, cela va modifier grandement notre manière d'appréhender les fonctions de $]0, +\infty[$ dans lui-même avec notre calculatrice.

Pour une introduction élémentaire aux nombres flottants et aux nombreuses conséquences qui dépassent le cadre de cette article, on ne saurait trop conseiller le premier chapitre de [1, 6] et le fameux article [2]. Le point de vue original de ce papier est, d'une part, d'aborder ces notions avec des outils du Collège : pourcentages, et du Lycée : logarithme, méthode des rectangles, et, d'autre part, d'introduire des distances relatives. Les distances relatives permettent de modéliser

(*) IUFM et Université de Nice, Laboratoire J. A. Dieudonné, UMR CNRS 6621.

la lecture et l'interprétation des résultats numériques d'une calculatrice. Elles permettent aussi de mieux comprendre la notion de précision en jeu pour les calculs de nos machines.

On va commencer par faire des tests numériques avec la calculatrice dans la partie 2. Cela nous conduira naturellement aux nombres flottants et aux écarts relatifs dans la partie 3. On fera le lien entre les écarts relatifs et les pourcentages dans la partie 4. Dans la partie 5, on s'intéressera tout particulièrement à l'inégalité triangulaire pour ces écarts relatifs. D'une part, on aura des relations intéressantes sur le lien entre augmentations ou diminutions répétées et inégalités triangulaires relatives. D'autre part, en itérant l'inégalité triangulaire, on aboutira naturellement via la méthode des rectangles à la distance logarithmique **dlog**. Cette dernière nous permettra, dans la partie 6, de mieux comprendre le comportement des fonctions puissances, exponentielles et périodiques de $]0, +\infty[$ dans lui-même lors de l'utilisation de nos calculatrices. Et aussi, l'on obtiendra le très important théorème des accroissements finis relatifs. On rassurera en partie le lecteur sur sa représentation usuelle des nombres dans la partie 7.

2. À vos calculatrices

On considère les deux nombres suivants $a := 2{,}003$, $b := 2,\ 004$. Ces deux nombres ont les trois premiers chiffres égaux et le quatrième diffère d'une unité. On va utiliser diverses fonctions f de notre calculatrice pour voir dans quelle mesure on conserve l'égalité des trois premiers chiffres. Quand a t-on les trois premiers chiffres de $f(a)$ et $f(b)$ égaux? On notera **NdCS** le nombre de chiffres conservés.

$f(x) =$	$f(a) \approx$	$f(b) \approx$	NdCS	$\lvert f(b) - f(a)\rvert \approx$
x	2,003	2,004	3	10^{-3}
$1\,000 \times x$	2003	2004	3	1
$8 \times x$	16,024	16,032	3	10^{-2}
$1\,234\,567 \times x$	2 472 837,7…	2 474 072,2…	3	$2 \times 10^{+4}$
$1/x$	4,992 5… 10^{-1}	4,990 02… 10^{-1}	3	3×10^{-4}
$1/(2\,004 \times \mathrm{x})$	2,491 2… 10^{-4}	2,490 03… 10^{-4}	3	10^{-7}
$\sqrt{x}$	1,415 2…	1,415 6…	4	4×10^{-4}
x^2	4,012 0…	4,016 0…	3	4×10^{-3}
x^{10}	1 039,46…	1 044,665	2	5
x^{100}	1,47… 10^{30}	1,54… 10^{30}	1	7×10^{28}
$\ln(x)$	6,946 4… 10^{-1}	6,951… 10^{-1}	2	5×10^{-4}
$\exp(x)$	7,411 25	7,418 6	3	7×10^{-3}
$\exp(50 \times x)$	3,123… 10^{43}	3,283… 10^{43}	1	$1{,}6 \times 10^{+42}$
$\exp(\exp(\exp(x)))$	3,48… 10^{718}	7,75… 10^{723}	0	$7 \times 10^{+723}$
$\exp(-50 \times x)$	3,201… 10^{-44}	3,045… 10^{-44}	1	$1{,}6 \times 10^{-45}$

Faites d'autres exemples. On remarquera l'étonnante stabilité de l'égalité de ces trois premiers chiffres, sauf pour quelques exemples dont nous parlerons plus loin. En revanche l'écart absolu $|f(b) - f(a)|$ est très variable.

Ainsi la précision la mieux conservée est celle du **nombre de chiffres significatifs**. Dans les exemples précédents l'on conserve souvent 3 chiffres significatifs. Mais cette notion de nombre de chiffres significatifs fait référence à un autre représentation de nombre que l'écriture décimale d'un nombre, elle fait référence aux nombres flottants.

3. Nombres flottants positifs

Rappelons la représentation scientifique des nombres utilisée par votre calculatrice.

On découpe $]0, +\infty[$ suivant les exposants positifs et négatifs de 10 :

$$]0,+\infty[= \cdots \cup [0,1;1[\cup [1;10[\cup [10;100[\cup \cdots = \bigcup_{p\in\mathbf{Z}} [10^p ; 10^{p+1}[.$$

Ainsi, soit $x > 0$, il existe un unique entier relatif p tel que $10^p \leq x < 10^{p+1}$. L'entier p s'appelle **l'exposant**. Et le nombre réel $m := \dfrac{x}{10^p}$ s'appelle **la mantisse**, $1 \leq m < 10$.

Notez que, pour un nombre strictement négatif x, alors x admet une écriture à virgule flottante. Il suffit de rajouter le signe : $x = -m\ 10^p$. En revanche zéro est une exception. Il n'admet pas d'écriture à virgule flottante. Ainsi zéro est représenté autrement dans la machine. On appelle cette représentation singulière de zéro, le zéro machine ou le zéro numérique.

Revenons aux nombres strictement positifs. Tout nombre réel strictement positif x se représente de manière unique sous la forme

$$x = m \times 10^p = m_1, m_2 m_3 \cdots m_N \cdots \times 10^p, \quad p \in \mathbf{Z}, \quad m \in [1;10[\qquad (1)$$

En fait, la machine calcule en base 2 et affiche le résultat en base 10. Mais, comme l'on s'en rendra compte plus loin, cela ne change rien à notre propos ici.

Pourquoi la machine à calculer utilise-t-elle cette représentation des nombres ? Imaginons que la machine ne puisse stocker que dix chiffres pour représenter un nombre $x > 0$. Une représentation habituelle de x est son développement décimal

$$x = x_4 x_3 x_2 x_1 x_0, x_{-1} x_{-2} x_{-3} x_{-4} x_{-5} = \sum_{k=-5}^{4} x_k 10^k \quad \text{où } x_k \in \{0,1,2,\ldots,8,9\} \qquad (2)$$

Alors un tel nombre ne pourra être que dans l'intervalle $[10^{-5}, 10[$. En revanche, toujours avec la même capacité mémoire d'une machine, si la représentation de x est faite à l'aide d'un nombre flottant, elle nous donne :

$$x = m_1, m_2 m_3 m_4 m_5 m_6 m_7 \times 10^{s \times p_1 p_2} \quad \text{avec } m_j, p_k \in \{0,1,2,\ldots,8,9\} \text{ et } s = \pm 1.$$

Cette fois-ci, x varie dans $[10^{-99}, 10^{100}[$. Ainsi, grâce aux nombres flottants de la calculatrice, on peut calculer avec des nombres beaucoup plus grands et des nombres beaucoup plus petits qu'en utilisant, avec la même capacité mémoire, la représentation décimale usuelle (2). D'une certaine manière, les nombres flottants de

la machine nous permettent d'approcher précisément l'infiniment grand et l'infiniment petit.

Revenons aux exemples de la partie 2. $a = 2{,}003$ et $b = 2{,}004$ ont leurs exposants nuls. Ils sont donc égaux à leurs mantisses. En revanche,

$$1\,000\,a = 2\,003 = 2{,}003 \times 10^{+3}$$

a 3 pour exposant et a pour mantisse.

Ainsi, comparer les premiers chiffres de $f(a)$ et $f(b)$ revient à comparer leurs mantisses. Lorsque deux nombres $x = m10^p$ et $x' = m'10^{p'}$ ont même exposant $p = p'$, on estimera $\Delta m = m - m'$. On dira par abus de langage que x' nous donne $k + 1$ chiffres significatifs de x si $p = p'$ et $|\Delta m| < 10^{-k}$. Incidemment, l'on voit que la notion de nombre de chiffres significatifs est invariante par multiplication ou division par les puissances de 10. Ceci est très différent de notre notion d'écart en valeur absolue dans le cadre de notre distance euclidienne.

En remarquant que $\Delta x = \Delta m \times 10^p$ et que $10^p \leq x < 10^{p+1}$, on a :

$$\frac{\Delta m}{10} < \frac{\Delta x}{x} \leq \Delta m \qquad (3)$$

Il en résulte : $\dfrac{\Delta x}{x} \leq \Delta m < 10\dfrac{\Delta x}{x}$. Ainsi $\dfrac{\Delta x}{x}$ représente assez fidèlement la précision que l'on a de la mantisse. De plus, pour des mantisses proches de 1 ou 10, $\dfrac{\Delta x}{x}$ représente mieux la notion de précision entre x et x' que Δm. Par exemple, si $x = 1{,}000$ et $x' = 0{,}999 = 9{,}99 \times 10^{-1}$, les mantisses sont totalement différentes, les exposants sont également différents. Et pourtant, x' est une excellente approximation de x. Effectivement on a bien $\dfrac{|\Delta x|}{x} \approx 10^{-3}$. Ce qui nous conduit naturellement à la notion d'écart relatif.

Notez de plus que $\dfrac{\Delta x}{x}$ convient aussi bien pour la représentation en base 10 de la mantisse (à l'affichage sur la calculatrice) que pour la base 2 (pour les calculs internes de la machine). En fait $\dfrac{\Delta x}{x}$ est *invariant par tout changement d'échelles* $\lambda > 0$: $\dfrac{\Delta \lambda x}{\lambda x} = \dfrac{\Delta x}{x}$.

4. Écarts relatifs, taux d'augmentation et de diminution

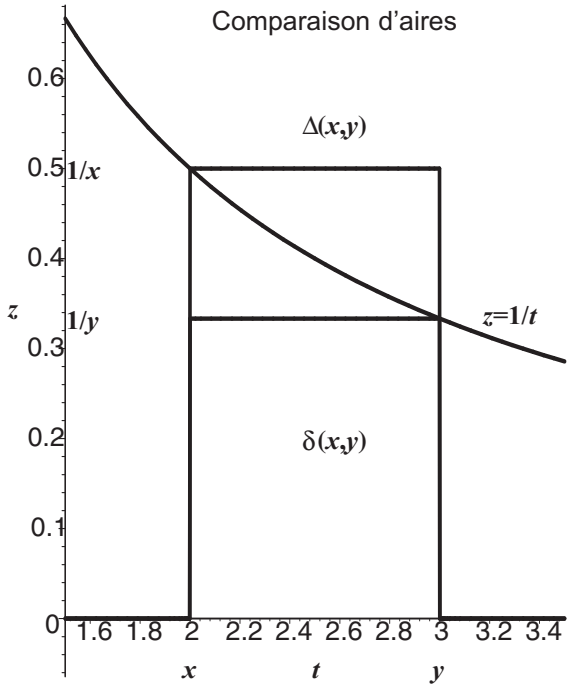
Soit x et y deux nombres strictement positifs, en prenant la valeur absolue de leur différence, on a naturellement deux écarts relatifs : $\dfrac{|x-y|}{x}$ et $\dfrac{|x-y|}{y}$. Considérons le plus petit et le le plus grand de ces écarts :

$$\delta(x,y) := \min\left(\frac{|x-y|}{x}, \frac{|x-y|}{y}\right) = \frac{|x-y|}{\max(x,y)}, \quad \Delta(x,y) := \max\left(\frac{|x-y|}{x}, \frac{|x-y|}{y}\right) \qquad (4)$$

Ainsi, on a toujours $\delta(x,y) \leq \dfrac{|x-y|}{x} \leq \Delta(x,y)$. Si l'on considère la courbe représentative de la fonction inverse : $t \mapsto \dfrac{1}{t}$ sur l'intervalle $[x, y]$, l'on s'aperçoit que $\delta(x,y)$ représente l'aire du plus grand rectangle vertical sous la courbe de la fonction inverse dans la demi-bande $[x, y] \times [0, +\infty[$. L'aire sous la courbe de la fonction inverse sera notée :

$$\text{dlog}(x,y) := \left|\ln\left(\frac{x}{y}\right)\right| = |\ln(x) - \ln(y)| \qquad (5)$$

De même, $\Delta(x, y)$ représente l'aire du rectangle vertical $[x, y] \times [0, +\infty[$. On obtient ainsi la figure suivante :



Ceci se traduit par les inégalités suivantes :

$$\delta(x,y) \leq \text{dlog}(x,y) \leq \Delta(x,y) \qquad (6)$$

On reviendra sur cette notion d'écart logarithmique dans la section suivante. Revenons aux deux écarts relatifs δ et Δ. On peut les interpréter en termes de taux d'augmentation ou de diminution. Pour fixer les idées, supposons que $x < y$, τ_A le taux d'augmentation et τ_D le taux de diminution :

$$x = y(1-\tau_D), \quad y = x(1+\tau_A) \quad \text{où } x < y \qquad (7)$$

Comme $x < y$, on a :

$$\delta(x,y) = 1 - \frac{x}{y} = \tau_D, \quad \Delta(x,y) = \frac{y}{x} - 1 = \tau_A.$$

Ainsi le plus petit écart relatif s'interprète comme le taux de diminution du plus grand au plus petit des nombres x et y, et le plus grand écart relatif comme le taux d'augmentation du plus petit au plus grand des deux nombres :

$$\delta(x,y) = \tau_D, \quad \Delta(x,y) = \tau_A.$$

On a bien que $\tau_D < \tau_A$. La différence stricte de ces deux taux est souvent méconnue de nos élèves, car en pratique (augmentations des prix, baisse du chômage, …) on a souvent affaire à des petits taux et on a bien $\tau_D \approx \tau_A$.

En multipliant les deux équations de (7), on obtient la relation entre le taux d'augmentation et le taux de diminution :

$$(1+\tau_A)(1-\tau_D) = 1 \qquad (8)$$

L'interprétation de cette égalité est très simple. On suppose toujours que $0 < x < y$. Si j'augmente x du taux τ_A, j'obtiens y. Si je diminue y du taux τ_D, j'obtiens x. Donc si j'augmente x du taux τ_A, puis si je diminue du taux τ_D, j'obtiens à nouveau x :

$$x(1+\tau_A)(1-\tau_D) = x.$$

Mais cette relation (8) entre les taux d'augmentation et de diminution est non linéaire. C'est ceci qui implique la différence des deux taux. En effet si on développe l'égalité (8) et si l'on simplifie, on obtient

$$\tau_A = \tau_D + \tau_A\tau_D \qquad (9)$$

On remarque aussi une autre dissymétrie entre les deux taux : τ_D est majoré par 1 alors que τ_A est non borné.

Revenons maintenant à notre problème de précision relative.

5. Distances relatives et distance logarithmique

Avant de rentrer dans les détails remarquez l'excellente corrélation entre le nombre de chiffres significatifs, l'écart relatif 6 et la distance logarithmique.

$f(x) =$	$f(a) \approx$	$f(b) \approx$	NdCS	$\delta(f(a),f(b))$	$\mathrm{dlog}(f(a),f(b))$
x	2,003	2,004	3	$0{,}5 \times 10^{-3}$	$0{,}5 \times 10^{-3}$
$1\,000 \times x$	2 003	2 004	3	$0{,}5 \times 10^{-3}$	$0{,}5 \times 10^{-3}$
$8 \times x$	16,024	16,032	3	$0{,}5 \times 10^{-3}$	$0{,}5 \times 10^{-3}$
$1\,234\,567 \times x$	2 472 837,7…	2 474 072,2…	3	$0{,}5 \times 10^{-3}$	$0{,}5 \times 10^{-3}$
$1/x$	4,992… 10^{-1}	4,990… 10^{-1}	3	$0{,}5 \times 10^{-3}$	$0{,}5 \times 10^{-3}$
$1/(2\,004 \times x)$	2,491…10^{-4}	2,490…10^{-4}	3	$0{,}5 \times 10^{-3}$	$0{,}5 \times 10^{-3}$
$\sqrt{x}$	1,415 2…	1,415 6…	4	$0{,}25 \times 10^{-3}$	$0{,}25 \times 10^{-3}$
x^2	4,012 0…	4,016 0…	3	10^{-3}	10^{-3}
x^{10}	1 039,46…	1 044,665…	2	$0{,}5 \times 10^{-2}$	$0{,}5 \times 10^{-2}$

x^{100}	$1{,}47\ldots\ 10^{30}$	$1{,}54\ldots\ 10^{30}$	1	$0{,}5\times10^{-1}$	$0{,}5\times10^{-1}$
$\ln(x)$	$6{,}946\,4\ldots\ 10^{-1}$	$6{,}951\ldots\ 10^{-1}$	2	$0{,}7\times10^{-3}$	$0{,}7\times10^{-3}$
$\exp(x)$	$7{,}411\,25\ldots$	$7{,}418\,6\ldots$	3	10^{-3}	10^{-3}
$\exp(50\times x)$	$3{,}123\ldots\ 10^{43}$	$3{,}283\ldots\ 10^{43}$	1	$0{,}5\times10^{-1}$	$0{,}5\times10^{-1}$
$\exp(\exp(\exp(x)))$	$3{,}48\ldots\ 10^{718}$	$7{,}75\ldots\ 10^{723}$	0	99,99%	12,3
$\exp(-50\times x)$	$3{,}201\ldots\ 10^{-44}$	$3{,}045\ldots\ 10^{-44}$	1	$0{,}5\times10^{-1}$	$0{,}5\times10^{-1}$

Ainsi l'on voit bien que le nombre de chiffres significatifs exacts donné par la machine correspond à cet écart relatif ou à l'écart logarithmique, sauf si le nombre de chiffres significatifs est trop mauvais : NdCS = 0.

On veut maintenant une distance relative d sur $]0, \infty[$. Une distance doit être symétrique : $d(x, y) = d(y, x)$ pour tout $x, y \in]0, +\infty[$. L'écart relatif $\dfrac{|\Delta x|}{x}$ n'est pas une expression symétrique en x et y. En revanche δ, Δ et dlog le sont.

On demande de vérifier l'axiome de séparation : la distance entre deux nombres n'est nulle que si ces deux points sont égaux. δ, Δ et dlog le vérifient :

$$\delta(x, y) = 0 \Leftrightarrow x = y.$$

D'autre part, l'on doit vérifier la célébrissime inégalité triangulaire :

pour tout $x, y, z \in]0, +\infty[$, $d(x, z) \leq d(x, y) + d(y, z)$.

On peux toujours supposer par symétrie que $x < z$ pour vérifier cette inégalité. Pour y, deux cas se présentent : ou bien $y \notin]x, z[$, ou bien $y \in]x, z[$. Le premier cas $y \notin]x, z[$ est le moins intéressant : δ, dlog et Δ vérifient simplement l'inégalité triangulaire dans ce cas. Par exemple, si $y > z$, on vérifie aisément que $\delta(x, z) < \delta(x, y)$.

En revanche le deuxième cas $y \in]x, z[$ est bien plus intéressant. En effet, si $x < y < z$, alors on va montrer à l'aide de calculs de pourcentages pour δ et Δ que $\Delta(x, z) > \Delta(x, y) + \Delta(y, z)$ et $\delta(x, z) < \delta(x, y) + \delta(y, z)$, alors que l'on vérifie aisément que $\text{dlog}(x, z) = \text{dlog}(x, y) + \text{dlog}(y, z)$.

Par exemple, pour passer de 1 à 3 on a fait une augmentation de 200%. Pour passer de 1 à 2 on a fait une augmentation de 100%, et de 2 à 3 de 50%. Finalement, on a $\Delta(1, 3) = 200\% > 150\% = \Delta(1, 2) + \Delta(2,3)$. De même, si on augmente deux fois de suite un nombre de 100% on l'a en fait augmenté de 400% !

Raisonnons en terme de taux d'augmentation pour bien nous convaincre que $\Delta = \tau_A$ n'est pas une métrique. Soit $\tau_{A1} > 0$ le premier taux d'augmentation pour passer de x à y : $\tau_{A1} = \Delta(x, y)$, $\tau_{A2} > 0$ le deuxième taux d'augmentation pour passer de y à z : $\tau_{A2} = \Delta(x, y)$, et τ_A le taux d'augmentation pour passer de x à z : $\tau_A = \Delta(x, y)$. L'inégalité triangulaire serait vraie pour Δ si on avait toujours $\tau_A \leq \tau_{A1} + \tau_{A2}$. Or, cette dernière inégalité est toujours fausse (sauf si un des deux taux intermédiaires est nul, ce qui est un cas sans intérêt). En effet, on a la relation suivante entre les taux :

$$(1+\tau_{A1})(1+\tau_{A2})=(1+\tau_A),$$

ce qui nous donne en développant et en simplifiant par 1 :

$$\tau_A=\tau_{A1}+\tau_{A2}+\tau_{A1}\tau_{A2}.$$

Ainsi, on a toujours $\tau_A>\tau_{A1}+\tau_{A2}$ dés que τ_{A1} et τ_{A2} sont non nuls : le taux d'augmentation ne vérifie pas l'inégalité triangulaire, donc Δ n'est malheureusement pas une métrique.

Maintenant, passons au taux de diminution entre x et z : $\tau_D=\delta(x,z)$, et $\tau_{D1}=\delta(x,y)$, $\tau_{D2}=\delta(y,z)$. De la relation $(1-\tau_{D1})(1-\tau_{D2})=1-\tau_D$, on tire :

$$\tau_D=\tau_{D1}+\tau_{D2}-\tau_{D1}\tau_{D2}.$$

Ainsi, on a toujours :

$$\tau_D<\tau_{D1}+\tau_{D2} \qquad (10)$$

dès que τ_{D1} et τ_{D2} sont non nuls. Par exemple, si l'on baisse deux fois un prix de 50%, on a finalement fait une réduction de 75% par rapport au prix initial.

On vient donc de démontrer la

Proposition : le taux de diminution est une métrique sur]0, +∞[.

δ *est une métrique sur* $]0,+\infty[$, *où* $\delta(x,y):=\min\left(\dfrac{|x-y|}{x},\dfrac{|x-y|}{y}\right)$.

Cependant, quelque chose trouble notre sens commun, l'inégalité *stricte* (10). En effet, si y est un point du segment $]x, z[$, on pense souvent que la distance de x à z devrait être la somme de celle de x à y avec celle de y à z :

$$0<x<y<z\Rightarrow d(x,y)+d(y,z)=d(x,z) \qquad (11)$$

Or, il n'en est rien à cause de l'inégalité *stricte* (10). Ceci peut être même pire : plus on prend de points intermédiaires entre x et y, plus la somme des longueurs intermédiaires augmente. Poussons le raisonnement jusqu'à son extrême limite. Pour cela, prenons une subdivision régulièrement espacée de l'intervalle $[x, z]$. Soit $x=x_0<x_1<\ldots<x_{n-1}<x_n=z$ avec $x_i-x_{i-1}=h=(z-x)/n$ pour $0<i\leq n$. Alors, $\delta(x_0,x_1)+\delta(x_1,x_2)+\ldots+\delta(x_{n-1},x_n)$ est en fait une somme d'aire de rectangles pour évaluer $\displaystyle\int_x^z\frac{\mathrm{d}t}{t}=\ln\left(\frac{z}{x}\right)$. Ainsi,

$$\delta(x,z)<\lim_{n\to+\infty}\sum_{j=0}^{n-1}\delta(x_j,x_{j+1})=\mathrm{dlog}(x,z).$$

On obtient donc, grâce à la méthode des rectangles, la distance logarithmique dlog. De plus, dlog vérifie toujours l'implication (11). Notre intuition géométrique en devient réconfortée. On peut d'ailleurs démontrer, [3], pour un sens mathématique raisonnable, que dlog est la seule métrique mesurant des écarts relatifs qui vérifie (11).

6. Théorème des accroissements finis relatifs

On veut mesurer la précision d'un résultat $f(x)$ lorsque l'on a seulement une approximation y de x. En fait, avec la machine, on évaluera $f(y)$. Et l'on voudrait bien un lien simple entre l'écart de $f(x)$ à $f(y)$ et l'écart de x à y. Pour cela, on voudrait utiliser une métrique censée représenter l'écart relatif $\frac{|\Delta x|}{x}$. On a le choix entre le taux de diminution δ ou la distance logarithmique dlog. Pour des erreurs de moins de 10% (au moins un chiffre significatif), on a des résultats semblables pour les deux métriques. On l'avait déjà remarqué dans le tableau d'exemples de la partie 5. On peut même démontrer de plus que :

$$\delta \leq 0,1 = 10\% \Rightarrow \delta \leq \text{dlog} \leq 1,054 \times \delta.$$

En effet, pour $x < y$, $\delta(x,y) = 1 - \frac{x}{y}$ et $\text{dlog}(x,y) = -\ln\left(\frac{x}{y}\right) = -\ln\big(1-\delta(x,y)\big)$, ainsi on a toujours $\text{dlog} = -\ln(1-\delta)$. Or, une étude de fonction (ou un développement en série entière) montre que $-\frac{\ln(1-u)}{u}$ est croissante pour $u \in]0, +\infty[$ et ainsi, pour $\delta < 0,1$, $\frac{\text{dlog}}{\delta} \leq -\ln(0,9) \times 10 \leq 1,054$.

Or, dlog a beaucoup d'avantages : elle va hériter des propriétés du logarithme, elle est plus simple à manipuler que δ qui est un minimum, elle vérifie (11), elle majore δ et on a des résultats optimaux et généraux. Pour des résultats plus précis et des comparaisons de dlog et δ, on renvoie à [3].

Par exemple, pour $f(x) = x^2$, on va apprécier le très bon comportement de la machine à calculer. Or, l'on apprend souvent que pour évaluer le carré d'un nombre précisément, plus ce nombre est grand, plus l'approximation de ce nombre doit être précise. On justifie ce raisonnement par le théorème des accroissements finis ou, plus simplement, par le calcul élémentaire suivant :

$$y = x + h \Rightarrow y^2 = x^2 + 2xh + h^2.$$

En pratique, on pense souvent que $|h| < 0,1$. On néglige à raison le terme en h^2 et l'on voit que la source principale de l'erreur que l'on commet en approchant x^2 par y^2 provient du terme $2hx$. Donc le terme principal de l'erreur est d'autant plus grand que x est plus grand.

Que dit notre calculatrice avec $h = 10^{-2}$?

x	$x^2 \approx$	$(x+h)^2 \approx$	NdCS	$\lvert\text{dlog}(x^2, (x+h)^2)\rvert \approx$
1	1	1,020 1	2	2×10^{-2}
100	10 000	10 002,000 1	4	2×10^{-4}
10^4	10^8	100 000 200,0	6	2×10^{-6}

Alors que l'écart absolu augmente de plus en plus, comme l'a prédit le développement de $(x + h)^2$, la machine nous offre de plus en plus de chiffres significatifs exacts. Ainsi, pour évaluer précisément les premiers chiffres significatifs d'un carré, il est inutile d'être très précis sur ce nombre. Par exemple, pour $123\,456\,789^2 = 15\,241\,578\,750\,190\,521$ et $123\,400\,000^2 = 15\,227\,560\,000\,000\,000$, on a les 4 premiers chiffres en commun avant d'élever ces deux nombres au carré. Et, l'on conserve 3 chiffres significatifs alors que le nombre envisagé est très grand $x \approx 10^8$ et la précision est très mauvaise en écart absolu $|h| \approx 5 \times 10^4$. Ceci contredit ce que l'on apprend sur le carré des grands nombres à condition de penser en termes de nombre de chiffres significatifs ou d'erreurs relatives. Qu'en est-il pour l'écart logarithmique ?

$$\mathrm{dlog}\left(x^2, (x+h)^2\right) = 2 \times \mathrm{dlog}\left(x, (x+h)\right),$$

car le logarithme d'un carré d'un nombre est le double du logarithme de ce nombre. Ainsi, l'écart relatif est seulement multiplié par 2, donc on perd au plus un chiffre significatif.

De même pour $f(x) = x^\alpha$ pour $x > 0$, on a

$$\mathrm{dlog}(f(x), f(y)) = |\alpha| \times \mathrm{dlog}(x, y).$$

Ainsi pour $\alpha = \pm 10$ on perd un chiffre significatif, pour $\alpha = \pm 100$ on en perd deux. Ceci correspond bien aux exemples numériques de la premières partie. Pour $|\alpha| \leq 1$ on a au moins le même nombre de chiffres significatifs. Ce qui explique le très bon comportement de la fonction racine carrée. Pour $\alpha = -1$, on trouve que la fonction inverse est une isométrie !

En revanche

$$\mathrm{dlog}\left(\mathrm{e}^x, \mathrm{e}^y\right) = |x - y| = \max(x, y) \times \delta(x, y),$$

ce qui rend compte du très mauvais comportement à la calculatrice de la fonction exponentielle pour les très grands nombres. De plus, comme le passage à l'inverse est une isométrie des distances relatives, la fonction exponentielle se comporte aussi mal au voisinage de $-\infty$.

Pour l'étude des accroissements relatifs, on utilise la notion d'élasticité, bien connue des économistes et des utilisateurs des coordonnées log-log. On obtient naturellement cette notion au Lycée pour des petits accroissements relatifs.

Définition : Élasticité e

À $f \in \mathrm{C}^1(]0, +\infty[,]0, +\infty[)$, *on associe son élasticité* $\mathrm{e}[f]$ *définie par*:

$$\mathrm{e}[f](x) := x \times \frac{f'(x)}{f(x)} = x \times \left(\ln(f)\right)'(x). \qquad (12)$$

On notera aussi $|\mathrm{e}|[f] = |\mathrm{e}|f||$.

En fait, comme on va le voir dans le théorème suivant, la notion d'élasticité pour les écarts relatifs est toute aussi importante que la notion de dérivée pour les écarts absolus.

Théorème : Égalité des accroissements finis logarithmiques
Soient $0 < a < b$, $f \in \mathrm{C}^0([a, b],]0, +\infty[)$, *dérivable sur* $]a, b[$. *Alors il existe* $c \in]a, b[$ *tel que*

$$\mathrm{dlog}\big(f(a), f(b)\big) = |\,\mathrm{e}\,|[f](c)\,\mathrm{dlog}(a, b) \qquad (13)$$

Démonstration : Pour obtenir l'égalité (13), il suffit de montrer qu'il existe c entre a et b tel que

$$\ln\left(\frac{f(b)}{f(a)}\right) = c\,\frac{f'(c)}{f(c)}\ln\left(\frac{b}{a}\right).$$

On peut la démontrer classiquement à l'aide du théorème de Rolle. Soit

$$\varphi(t) = \big(\ln\big(f(t)\big) - \ln\big(f(a)\big)\big) - \mathrm{L}\big(\ln(t) - \ln(a)\big),$$

avec L tel que $\varphi(b) = 0$. Comme $\varphi(a)$ est aussi nulle, il existe $c \in]a, b[$ tel que $\varphi'(c) = 0$, ce qui nous donne la valeur de L escomptée.

On peut aussi voir cette égalité comme un cas particulier du théorème des accroissements finis généralisés, [5] p. 281 : $h'(c)(g(b) - g(a)) = g'(c)(h(b) - h(a))$ avec $h = \ln$ et $g = \ln \circ f$.

Par exemple, $\mathrm{e}[t \mapsto t^{\alpha}] = \alpha$ est constante, et $\mathrm{e}[\exp](x) = x$ est non bornée.

On a l'habitude d'être très confiant vis-à-vis des fonctions périodiques régulières. On a tort lorsqu'on utilise une calculatrice avec de grands nombres. En effet, considérons $f(x) = 2 + \sin(x)$. Le nombre 2 nous permet de rester parmi les nombres strictement positifs. $\mathrm{e}[f](x) = x \times \dfrac{\cos(x)}{2 + \sin(x)}$ est non bornée. On peut donc s'attendre à des erreurs pour les grandes valeurs de x. En effet, travaillons avec 10 chiffres significatifs, $2 + \sin(10^9\,\pi) = 2$ alors que la machine donne :

$$2 + \sin(3{,}141\,592\,654 \times 10^9) = 2{,}398\,798\,944\,8,$$

ce qui est grossièrement faux. On retiendra que les fonctions périodiques régulières se comportent très bien par rapport à l'écart absolu grâce à la classique inégalité des accroissements finis, mais qu'en général elles se comportent mal pour les grands nombres en utilisant une calculatrice.

7. Épilogue

Pour rassurer nos lecteurs, il faut savoir que les calculatrices utilisent plus de chiffres significatifs en calcul interne qu'à l'affichage (de 20% à 50% de plus !), ceci, pour surmonter en partie les problèmes rencontrés. D'ailleurs, cela permet souvent à l'utilisateur de conserver sa représentation décimale habituelle [2].

Cependant, en pratique, on connaît rarement 15 chiffres significatifs d'un nombre. À part π, $\sqrt{2}$, la constante de gravitation universelle et quelques autres nombres, on ne connaît tout au plus que quelques chiffres significatifs d'un nombre. En revanche, toujours en pratique, le choix de l'unité est crucial pour se ramener justement à l'unité et retrouver notre bonne vieille intuition euclidienne de la mesure

des distances. Par exemple, on n'évalue pas le PIB d'un pays en euros mais en milliards d'euros, le nombre de microbes lors d'une maladie à l'unité près mais en centaine de milliers, … On parle toujours en ordre de grandeur et non en grandeur absolue à partir d'une unité inappropriée au problème étudié. D'ailleurs, on retrouve que $\dfrac{\Delta x}{x} \approx x$ lorsque $x \approx 1$.

Pour une étude un peu plus poussée des distances relatives au niveau du CAPES de mathématiques, on propose au lecteur l'article [3].

Dans cet article, on a à peine parlé du zéro numérique et du passage des nombres positifs aux nombres négatifs dans la partie 3. Cependant, il faut savoir que cela entraîne des erreurs gravissimes remarquablement exposées dans [2], [1], [6]. Par exemple, la calculatrice peut se tromper grossièrement lors de la soustraction de nombres très très proches. Pour donner une idée au lecteur du danger et en suivant l'esprit des distances relatives, voici un avant-goût géométrique du *zéro numérique*. On a vu que le passage à l'inverse est toujours une isométrie car les distances relatives sont invariantes par changement d'échelles. Ainsi la distance d'un nombre strictement positif x à un nombre très petit $\varepsilon > 0$ est la même que celle de $1/x$ à $1/\varepsilon$! Autrement dit, 0 est aussi loin de x que $1/x$ l'est de $+\infty$! Plus précisément, on a même :

$$\lim_{\varepsilon \to 0} \delta(x, \varepsilon) = \lim_{\varepsilon \to 0} \delta\left(x, \frac{1}{\varepsilon}\right) = 1 = 100\%,$$

donc, en prolongeant $\delta(x, \cdot)$ en 0 et en $+\infty$, on a :

$$x > 0 \Rightarrow \delta(x, 0) = \delta(x, +\infty).$$

Mais ceci est une autre histoire qui vous sera prochainement racontée dans [4].

Références

[1] Demailly, *Analyse numérique et équations différentielles*, chapitre 1, PUG, (1991).

[2] G.E. Forsythe, *Pitfalls in computations, or why a math book is not enough*, Amer. Math. Monthly, 77 (1970), p. 931-956.

[3] Junca, *Nombres positifs, métriques et calculatrices*, http ://www.math.unice.fr/junca, preprint (2003), 15 p.

[4] Junca, *La calculatrice et le zéro numérique*, en préparation.

[5] Tissier & Mialet, *Analyse à une variable réelle*, Bréal, (2000).

[6] D. Schatzman Analyse numérique, chapitre 0, (1991).