STICHTING
MATHEMATISCH CENTRUM
2e BOERHAAVESTRAAT 49
AMSTERDAM

DR 13r

Erreurs d'arrondiment dans les calcules systematiques.

(Les machines a calculer et la pensee humaine,
C.N.R.S., Paris 1953, p 285-293).

A. van Wijngaarden.

1953

13

COLLOQUES INTERNATIONAUX
DU
CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE

XXXVII

Les machines à calculer
et la
pensée humaine

PARIS
8-13 janvier 1951

Extrait

Editions du Centre National de la Recherche Scientifique
13, Quai Anatole France — PARIS (7e)
1953

ERREURS D'ARRONDIMENT DANS LES CALCULS SYSTÉMATIQUES

par A. VAN WIJNGAARDEN (1)

Il est bien connu que dans les procédés d'itérations, utilisés dans les machines à calculer modernes à grande vitesse, les erreurs d'arrondiment peuvent avoir une influence sérieuse sur le résultat du calcul. Il est vrai que les calculs à la main n'en sont pas épargnés et ces erreurs peuvent même y être plus dangereuses, puisque l'on a tendance à garder un nombre de décimales peu élevé pour s'épargner du travail, et qu'il est plus aisé de doubler la précision des calculs avec une machine à grande vitesse. Mais, d'un autre côté, le nombre d'étapes de calcul est généralement beaucoup plus grand dans une machine à calculer, et, ceci, entre beaucoup d'autres raisons, parce que le mathématicien chargé du programme préfère des procédés simples à exécuter un certain nombre de fois plutôt que des procédés compliqués qui sont difficiles à programmer. En outre, dans des calculs à la main, on fait fréquemment usage de tables de fonctions, alors que dans une machine à calculer on peut préférer calculer des fonctions auxiliaires pendant le calcul principal pour réduire le volume de la mémoire mécanique. En particulier quand ces calculs auxiliaires consistent en la résolution de relations de récurrence ou d'équations différentielles, dont la solution suit le rythme du problème principal, l'arrondiment joue un rôle important.

Je me propose d'esquisser quelques traits assez essentiels de la structure des erreurs d'arrondiment dans les calculs d'itération, au moyen de l'analyse des procédés particuliers plutôt simples, dans laquelle on peut facilement reconnaître cette structure dans des calculs plus compliqués.

D'abord je donnerai rapidement une vue d'ensemble sur plusieurs types de procédés d'itérations. Ils sont tous semblables en

(1) Mathematisch Centrum, Amsterdam (Pays-Bas).

ce qu'un certain cycle de calcul se répète de nombreuses fois consécutivement. Ils peuvent être d'un type homogène ou non. En outre, les résultats d'un certain cycle sont repris ou non dans le cycle suivant. Ils ne s'y introduisent pas par des procédés comme la sous-tabulation ou la formation des différences d'une table donnée de fonction. Ici l'analyse des erreurs d'arrondiment est essentiellement celle de la détermination de la distribution de fréquence des erreurs de différentes grandeurs. Ils s'introduisent effectivement dans le cycle suivant par des procédés comme l'intégration, la résolution d'équations différentielles ou d'équations aux différences ou d'autres types d'équations fonctionnelles. L'intégration et la résolution d'équations différentielles ou aux différences non homogènes sont de type non homogène. Une information extérieure s'introduisant dans les calculs d'une façon continue. Dans les calculs de type homogène, la solution se développe indépendamment et doit probablement être plus sérieusement assujettie aux erreurs spécialement quand existent plusieurs solutions fondamentales de différents ordres de grandeur, lorsque le calcul est dit instable.

Certainement, le phénomène le plus remarquable dans les erreurs d'arrondiment est qu'elles paraissent être dues au hasard. Ceci crée un « bruit » dont le niveau augmente plus ou moins rapidement au cours du calcul. Dans les calculs instables, même le plus bas niveau de bruit peut éventuellement devenir désastreux. Nous retiendrons davantage notre attention sur les calculs stables. Même là, cependant, des erreurs plutôt graves peuvent survenir, dues aux effets de non-hasard de l'arrondiment. Deux espèces importantes sont dues, l'une à des erreurs sur les paramètres, qui jouent dans les calculs un rôle et qu'on appelle *erreurs de paramètres,* et l'autre à la « condensation » de la fonction continue dans une fonction qui prend seulement que des valeurs discontinues, ce qu'on pourrait appeler *erreurs de condensation.*

On peut d'abord prêter quelque attention au bruit. Nous pouvons donc étudier la fréquence des erreurs d'arrondiment dans des combinaisons linéaires de données arrondies, ceci étant un problème très important et, en outre, pas trop compliqué. Par une définition convenable de toutes les opérations numériques, les nombres arrondis peuvent être pris comme des nombres entiers. On désignera alors par Af la valeur arrondie d'une quantité et par Bf la partie fractionnaire, de façon que $f = Af + Bf$, $Af \equiv 0 \pmod 1$ et $|Bf| \leqslant \frac{1}{2}$. Dans les calculs à la main, il est d'usage courant et d'une très bonne pratique d'éviter l'ambiguité

par la définition plus subtile $Af \equiv 0 \pmod 1$ et $|Bf| < \frac{1}{2}$ ou $Af = 0 \pmod 2$ et $|Bf| = \frac{1}{2}$. Ceci rend en outre l'erreur d'arrondiment symétrique. Dans une machine à calculer on préfère presque toujours pour sa simplicité le système asymétrique d'arrondiment, défini par $Af \equiv 0 \pmod 1$ et $-\frac{1}{2} < Bf \leqslant \frac{1}{2}$. Nous négligerons ici cette petite variante, et supposerons qu'on utilise un système symétrique d'arrondiment. Etant donné une grande série de valeurs de la fonction f_k, le problème revient alors à déterminer $f = \sum_{k=1}^{n} a_k f_k$. Ce qu'on calcule actuellement est $g = \sum_{k=1}^{n} (Aa_k)(Af_k)$. L'erreur est donc $\psi = f - g$. Il peut se trouver aussi qu'on ait seulement à calculer Af. Le calcul donne alors Ag et l'erreur est $P = Af - Ag$. Si le calcul est répété de nombreuses fois sur des suites différentes de n termes, on peut définir les fréquences $v(\psi)$ et $\Omega(P)$ des erreurs. Sous la condition que Bf_k ait une distribution homogène entre $-\frac{1}{2}$ et $\frac{1}{2}$, et que tous les Bf_k soient indépendants, un calcul assez compliqué, nécessitant la transformation de Laplace et des considérations sur la théorie des nombres, fournira les expressions explicites de $v(\psi)$ et $\Omega(P)$. On peut également traiter le cas où Bf_k dépend des m valeurs précédentes. Nous avons appliqué cette théorie rigoureuse de la microstructure à des problèmes comme ceux des fluctuations dans les différences d'un ordre élevé de valeurs tabulées de fonctions et des erreurs dans la sous-tabulation.

Tous ces calculs sont toutefois si compliqués qu'il est très désirable de connaître dans quelles circonstances des méthodes d'approximation fourniront des résultats précis quand n est grand, ce qui est le cas le plus intéressant. Cela est possible parce que la théorie donne aussi des expressions asymptotiques. A des fins pratiques, il est important de connaître si la distribution est approximativement normale ou non. Quand les B_k sont indépendants et également répartis, cette distribution est alors asymptotiquement normale lorsque :

$$\lim_{n \to \infty} \left(\sum_{k=1}^{n} \left| a_k^3 \right| \right)^{\frac{1}{3}} \left(\sum_{k=1}^{n} \left| a_k^2 \right| \right)^{-\frac{1}{2}} = 0.$$

S'il en est ainsi, la déviation type σ de la distribution normale est donnée par $\sigma^2 = \frac{1}{12} \sum_{k=1}^{n} a_k^2$, et il est plus que probable que la

valeur du bruit n'excèdera pas un petit multiple de σ. Pour simplifier nous appellerons σ le niveau de bruit. Il n'est en rien évident que la distribution est asymptotiquement normale. Dans le cas de l'interpolation à un ordre élevé, il n'en est pas ainsi. On voit ici le danger de calcul trop peu soigné. Quoiqu'en général, des considérations superficielles fournissent des résultats raisonnables, elles peuvent cacher des traits essentiels de structure dans les autres cas. Supposons que nous effectuions une interpolation Lagrangienne d'un certain ordre, l'intervalle de la suite f_k étant fractionné en un nombre élevé de parties p bien déterminées, alors, dans des conditions qui sont généralement satisfaites, la probabilité $\Omega(0)$ pour qu'il n'y ait aucune erreur due à l'arrondiment est une fonction totalement tératologique de p, car elle est discontinue pour chaque valeur rationnelle de p. Pour $p = 0$ par exemple, nous avons visiblement $\Omega(0) = 1$, alors que pour p très petit, nous avons $\Omega(\theta) = \frac{3}{4}$.

Jetons maintenant un regard sur les erreurs de paramètre. Un exemple très simple est fourni par la relation de récurrence suivante :

$$f_n = e^{-h} f_{n-1}, \quad f_0 = 1$$

dont la solution est visiblement $f_n = e^{-nh}$ en l'absence d'arrondiment. Maintenant, partageons les effets d'arrondiment en deux parts en tenant d'abord compte de l'arrondiment du paramètre e^{-h}, mais en supposant en même temps que pour le reste l'arrondiment n'a pas lieu. L'erreur qui dans ce cas particulier est introduite dans la solution peut être appelée erreur de paramètre. Une autre équation est résolue, à savoir :

$$f'_n = (Ae^{-h}) f'_{n-1}, \qquad f'_0 = 1.$$

Posons maintenant :

$$Ae^{-h} = e^{-h'},$$

où :

$$|h' - h| \ll 1.$$

Soit r, la valeur numérique du dernier chiffre décimal ou binaire, et soit $|\vartheta| \leqslant \frac{1}{2}$; on a donc $e^{-h'} = e^{-h} + \vartheta r$. Pour $r \ll h \ll 1$ nous trouvons: $h' = h + \vartheta r$, et par conséquent la véritable solution, débarrassée du bruit, est

$$f'_n = e^{-n(h+\vartheta r)} \approx (1 - n\vartheta r)\, e^{-nh}.$$

L'erreur de paramètre est donc égale à $n\vartheta r e^{-nh}$, en admettant, naturellement que le calcul est effectué avec un point décimal ou

binaire fixe. En réalité, à cette solution il se superpose encore un bruit d'un certain niveau. Car on calcule en réalité :

$$f''_n = \mathrm{A}\left\{ (\mathrm{A}e^{-h})\, f_{n-1} \right\}, \qquad f''_0 = 1.$$

A cause de la linéarité de l'équation, ce nouveau système d'erreurs d'arrondiment est parfaitement décrit par un système d'erreurs ε_k, produit au $k^{\text{ième}}$ échelon et se développant ensuite indépendamment, suivant la même loi exponentielle que f'_n, à cela près, toutefois, que n doit être remplacé par $n - k$. Le bruit total au $n^{\text{ième}}$ échelon est donc $\sum_{k=1}^{n} e^{-kh'} \varepsilon_k$. Si nous négligeons la petite influence de la corrélation existante sur ε_k, le bruit est alors normalement distribué pour les grandes valeurs de n, et après quelques calculs, nous trouvons que le niveau de bruit doit être $\sigma \sim \dfrac{r}{\sqrt{24h}}$ donc constant.

Un exemple plus compliqué est fourni par une méthode propre à donner les sinus ou cosinus, à savoir les relations de récurrence

$$f_n = 2\cos h\, f_{n-1} - f_{n-2}, \qquad f_0 = 0, \qquad f_1 = \sin h,$$

d'où :

$$f_n = \sin nh,$$

$$g_n = 2\cos h\, g_{n-1} - g_{n-2}, \qquad g_0 = 0, \qquad g_1 = \cos h,$$

d'où :

$$g_n = \cos nh.$$

Les avantages de ce procédé sont que, soit le sinus soit le cosinus peuvent être calculés indépendamment, et qu'une multiplication seulement est nécessaire.

A nouveau considérons d'abord la solution sans bruit avec des erreurs de paramètres. Si nous posons $\mathrm{A}\,(2\cos h) = 2\cos h'$, nous avons pour $r \ll h \ll 1$, $h' = h\,(1 + \delta)$ avec $\delta = \dfrac{3r}{(2h^2)}$. La solution est alors, si, par souci de précision, on remplace nh par x :

$$f'_n = \frac{1}{1+\delta} \sin (1+\delta)\, x,$$

$$g'_n = \cos (1+\delta)\, x + h \sin (1+\delta)\, x.$$

La phase et l'amplitude sont affectées, mais l'une et l'autre sont

indépendantes de n. Quand, en outre, $\delta \ll 1$; et $x \ll \frac{1}{\delta}$, nous avons les résultats simples :

$$f'_n = \sin x - \delta (\sin x - x \cos x)$$
$$g'_n = \cos x - \delta x \sin x$$

L'erreur de paramètre est ici, par conséquent, en première approximation, une fonction additive, croissante avec x, et proportionnelle à δ.

Le bruit peut être analysé d'une manière simple comme dans l'exemple précédent. On a trouvé que :

$$\sigma^2 = \frac{r^2}{48} \frac{(2n+1) \sin h - \sin (2n+1) h}{\sin^3 h}$$

ce qui donne pour $x \ll 1$:

$$\sigma = \frac{n \sqrt{n}}{6} r,$$

et pour $x \gg 1$:

$$\sigma = \frac{\sqrt{n}}{2 \sqrt{6h}} r.$$

Le bruit augmente donc rapidement d'abord, mais beaucoup plus lentement ensuite.

Au titre de comparaison et aussi comme exemple plus compliqué, nous considèrerons un autre procédé de génération des sinus et cosinus, à savoir :

$$f_n = \cos h \, f_{n-1} + \sin h \, g_{n-1}, \qquad f_0 = 0,$$

d'où :

$$f_n = \sin nh,$$

$$g_n = - \sin h \, f_{n-1} + \cos h \, g_{n-1}, \qquad g_0 = 1,$$

d'où :

$$g_n = \cos nh.$$

Ici, le sinus et le cosinus doivent être calculés simultanément et, en outre, deux multiplications sont nécessaires pour chacun d'eux.

Pour déterminer la solution sans bruit avec erreur de paramètre, c'est-à-dire f'_n, g'_n, nous devons tenir compte que h a été mutilé deux fois, de telle sorte qu'on doit introduire deux nouveaux paramètres définis par $\cos h' = A \cos h$, et par $h'' = A \sin h$. Si nous posons $a^2 = \cos^2 h' + \sin^2 h''$, et $\operatorname{tg} H = \frac{\sin h''}{\cos h'}$, nous trou-

vons alors :

$$f'_n = a^n \sin nH$$
$$g'_n = a^n \cos nH.$$

Avec $\cos h' = \cos h + \vartheta' r$, $\sin h' = \sin h + \vartheta'' r$, $\delta' = \frac{1}{2}\vartheta' r$, et $\delta'' = \frac{1}{2}\vartheta'' r$, nous avons pour $r \ll h \ll 1$ et $nr \ll 1$:

$$f'_n = \sin x + \delta' x \sin x + \delta'' x \cos x,$$
$$g'_n = \cos x + \delta' x \cos x - \delta'' x \sin x.$$

On doit noter que δ' et δ'' sont, en dehors des effets de ϑ, ϑ' et ϑ'', de l'ordre de δh et par conséquent beaucoup plus petits que dans la méthode précédente. Pour x grand, toutefois, l'erreur de paramètre est beaucoup plus grave, parce que l'amplitude croît et décroît suivant une loi exponentielle.

Le niveau de bruit est de $\frac{\sqrt{n}}{2\sqrt{3}} r$, et il est donc beaucoup plus bas que dans la méthode précédente.

Dans tous les cas mentionnés, le bruit et les erreurs de paramètre sont proportionnels à r. Ceci peut sembler, à première vue, une propriété plutôt évidente, mais nous allons montrer que les erreurs de condensation ne suivent pas cette règle.

Nous prendrons d'abord un exemple très facile et familier, à savoir la solution d'une équation quadratique $ax^2 + bx + c = 0$, au moyen de la formule $2ax = -b \pm \sqrt{b^2 - 4ac}$. Naturellement, les erreurs d'arrondiment sur a, b et c, donnent une erreur sur x. Une analyse simple montre que, pour $4ac \ll b^2$, l'erreur relative sur x est approximativement deux fois celle des coefficients. Mais si nous effectuons réellement le calcul au moyen de cette formule spécifique, un groupe de chiffres significatifs en fait disparaître un autre, et il en résultera visiblement une grande erreur sur x. La solution de ce paradoxe est assez simple : le reste des chiffres provient des chiffres non significatifs du calcul. C'est seulement dans la formation de $b^2 - 4ac$ qu'aucun chiffre n'est perdu et c'est de ce nombre « de précision double » qu'on doit extraire la racine carrée. Un tel changement temporaire dans la précision est, toutefois, assez inquiétante. Naturellement, dans ce cas particulier, il y a de meilleurs procédés pour trouver x.

La situation est ici autre que celle où l'erreur de paramètre était acceptable. Là, dès l'origine, la solution du problème n'était pas mieux déterminée par la précision connue des données. Ici, par contre, la solution est très bien définie mais l'information

est brusquement perdue en un certain point des calculs, parce qu'un nombre est réduit à un degré tel qu'il ne contient plus l'information nécessaire.

Nous donnerons maintenant un exemple plus intéressant, à savoir un procédé utilisé dans les machines à calculer à grande vitesse pour trouver les fonctions trigonométriques inverses. Soit par exemple x défini par $\cos x = a$. Le procédé d'itération $\cos 2^n x = -1 + 2\cos^2 2^{n-1}x$, donnera alors une suite de nombres. On peut, d'après les signes de ces nombres, construire une fraction binaire qui donnera x, après multiplication par π. Comme le nombre d'itérations est faible, à savoir 30 ou 40, le niveau de bruit est bas, et les erreurs de paramètres ne se présentent pas, puisque les paramètres donnés sont exacts. Cependant, il y a une sérieuse difficulté. En fait, il est évident que lorsque x est voisin de 0 ou de π, il est mal défini par la valeur arrondie de a. Cela n'a pas d'importance en soi, simplement, parce que le problème n'est pas mieux défini. Mais si x est voisin de $\frac{1}{2}\pi$, il est très bien défini par a. Cependant, le tout premier échelon du procédé fournit le cosinus de $2x$ qui est voisin de π, et là est la complication. D'autre part, si x est voisin de $\frac{1}{4}\pi$ ou $\frac{3}{4}\pi$, la situation est dangereuse après deux étapes du procédé, et ainsi de suite. Si nous étudions le phénomène de plus près nous arrivons alors au résultat suivant.

Soit $\mathrm{R}^2 = \frac{1}{4}r$ et $x = \frac{1}{2}\pi - u$, $\frac{\varepsilon}{\mathrm{R}}$ dépend alors de $\frac{u}{\mathrm{R}}$ de la façon suivante :

$$\frac{\varepsilon}{\mathrm{R}} = \frac{u}{\mathrm{R}} - \sqrt{2k}, \text{ pour } \sqrt{2k-1} < \frac{u}{\mathrm{R}} < \sqrt{2k+1}$$

$$(k = \ldots, -2, -1, 0, 1, 2, \ldots).$$

Il y a, par conséquent, une petite région, la « zone de danger » autour de $x = \frac{1}{2}\pi$, où l'erreur excède considérablement le niveau de bruit. Autour de $x = \frac{1}{4}\pi$ et $x = \frac{3}{4}\pi$, une zone de danger semblable existe, mais deux fois plus petite et où les erreurs sont aussi deux fois plus petites, et ainsi de suite. Pour se faire une idée de la largeur de la zone de danger, nous la limiterons de telle sorte qu'à l'extérieur de ses limites, l'erreur soit inférieure à kr. La largeur autour de $x = \frac{1}{2}\pi$, sera alors de $\frac{1}{hk}$. Si k est pris égal à 10, la largeur totale de toutes les zones de danger est de l'ordre de 1 % de l'intervalle total de x. La plus grande erreur est visiblement R, et par conséquent la moitié du nombre de déci-

males peut être perdu au cours du calcul. Si c'est *arc tg* x qui est calculé au lieu de *arc cos* x et si le point arithmétique est flottant, la difficulté ne se présente pas.

Nous ne nous arrêterons pas au cas du calcul instable où le niveau de bruit tend à monter si vite que les effets dus aux erreurs d'arrondiment submergent complètement la solution recherchée. Au contraire, je montrerai sur un dernier exemple que dans un cas où on s'attendrait à ce que cela se produise, une remarquable coïncidence sauve le calcul. Posons le problème de la résolution numérique de l'équation différentielle $y'' = y$, avec les conditions initiales $y(0) = -y'(0) = 1$. La solution recherchée est par conséquent $y = e^{-x}$. Nous adopterons, pour la résolution, la méthode dite des séries de Taylor. Comme, néanmoins $f^{(2k)} = f$ et $f^{(2k+1)} = f'$, il sera seulement nécessaire de calculer y et y'. L'échelon de calcul peut être variable. Nous trouvons alors que le calcul tombe brusquement à la solution des relations de récurrence :

$$f(x+h) = ch\,h\,f(x) + sh\,h\,f'(x),$$
$$f'(x+h) = sh\,h\,f(x) + ch\,h\,f'(x).$$

Les deux solutions fondamentales de l'équation différentielle, à savoir e^x et e^{-x} sont solution de ces relations. Après que les calculs ont été poursuivis quelque temps, la solution est entachée d'erreurs d'arrondiment. Mathématiquement parlant, la solution est alors une combinaison linéaire des deux solutions fondamentales, et comme la solution parasite, à savoir e^x, augmente très vite par rapport à e^{-x}, on doit craindre que l'erreur ne grandisse très vite. Si, toutefois, f et f' sont donnés avec le même nombre de chiffres, cela n'arrivera pas. Car si $f(x)$ est exactement égal à $-f'(x)$, cette égalité parfaite est vraie aussi pour $x + h$, parce que les opérations numériques se font pour deux fonctions alors identiques. Comme les conditions initiales satisfont à cette égalité, ceci sera toujours vrai. Soit, à une certaine étape, $y(x) = -y'(x) = (1+\delta)\,e^{-x} + \varepsilon e^{x}$. Avant d'arrondir : $y(x+h) = -y'(x+h) = (1+\delta)\,e^{-(x+h)} + \varepsilon e^{-ch}\,e^{(x+h)}$. Par conséquent, chaque erreur est supprimée immédiatement, et la solution n'aura aucune chance de se développer elle-même dans le bruit. Mais l'argumentation ne vaut pas si y et $-y'$ ne sont pas exactement égaux. Si, par conséquent, nous essayons d'améliorer la solution en donnant plus de décimales à y, mais moins de décimales supplémentaires à y' cette métastabilité remarquable s'évanouit et la précision de la solution disparaît immédiatement !