Bulletin de la Classe des sciences

Le calcul numérique des fonctions élémentaires dans la machine mathématique IRSIA-FNRS

V. Belevitch, F. Storrer

Résumé
Après quelques généralités sur la précision des calculs dans les machines à virgule flottante, on décrit diverses méthodes mathématiques, dont certaines originales, pour l'approximation des fonctions par des polynômes, avec une erreur relative maximum prescrite. On décrit le calcul des fonctions x-1, x-1|2, sin x, arc tg x, 10x, 10x - 1, log10x et log10(1 + x) dans la machine belge IRSIA-FNRS, pour une erreur relative de l'ordre de 10-14, et on établit les valeurs numériques des constantes qui y interviennent. Pour chaque fonction on donne une discussion détaillée des erreurs et du procédé de contrôle.

Le calcul numérique des fonctions élémentaires dans la machine mathématique IRSIA-FNRS

par V. BELEVITCH et F. STORRER (*)

Résumé. — Après quelques généralités sur la précision des calculs dans les machines à virgule flottante, on décrit diverses méthodes mathématiques, dont certaines originales, pour l'approximation des fonctions par des polynômes, avec une erreur relative maximum prescrite. On décrit le calcul des fonctions x^{-1}, $x^{-1/2}$, $\sin x$, arc tg x, 10^x, $10^x - 1$, $\log_{10}x$ et $\log_{10}(1 + x)$ dans la machine belge IRSIA-FNRS, pour une erreur relative de l'ordre de 10^{-14}, et on établit les valeurs numériques des constantes qui y interviennent. Pour chaque fonction on donne une discussion détaillée des erreurs et du procédé de contrôle.

1. Calculs a virgule flottante.

On sait que dans les machines à virgule flottante, un nombre est représenté par une mantisse et un exposant séparés, une notation telle que

$$+\ 314159\ldots\ +\ 01 \tag{1}$$

signifiant

$$+\ 0{,}314159\ldots\ 10^{+01}.$$

La virgule est toujours sous-entendue en tête de la mantisse, et le premier chiffre de celle-ci ne peut être un zéro : un résultat de calcul tel que $0{,}004\ldots\ 10^{00}$ est en effet normalement réaligné par la machine elle-même en $0{,}4\ldots\ 10^{-02}$. Dans la machine IRSIA-FNRS, la mantisse a quinze chiffres significatifs, et l'exposant en a deux, d'ailleurs limités à 49, de façon qu'il n'y ait

(*) Présentés par M. Manneback.

jamais report au troisième rang lors d'une addition de deux exposants (ce qui se produirait pour 50 + 50) ; dans ce but, un résultat dont l'exposant dépasse 49 en module provoque une alarme.

Les avantages de la notation à virgule flottante sont bien connus : quel que soit son ordre de grandeur (dans le champ de variation admis par l'exposant), un nombre peut être représenté avec une erreur inférieure à une unité du quinzième rang (1) ; dans le cas le plus défavorable où la mantisse est 0,1, ceci correspond à une erreur relative de 10^{-14}. D'autre part, pour le calcul de l'inverse p. ex., il suffit de changer le signe de l'exposant, et d'inverser la mantisse au moyen d'un procédé d'approximation dont la validité peut être restreinte a priori à un intervalle d'une décade, s'étendant de 0,1 à 1.

2. Les fonctions élémentaires.

L'unité arithmétique de la machine n'est capable que de combiner des nombres (compte-tenu de leurs exposants) par addition, soustraction et multiplication. Elle est également à même d'effectuer certaines *altérations* simples, p. ex. celle qui consiste à changer un exposant en mantisse (ceci intervient dans le calcul du logarithme), changer le signe d'un nombre ou de son exposant (exemple de la division ci-dessus), etc. Les autres fonctions arithmétiques (inverse, racine carrée, etc.), et a fortiori les fonctions transcendantes usuelles, se calculent au moyen de programmes établis une fois pour toutes, et ne pouvant comporter que des opérations arithmétiques ou des altérations du type prévu. Il s'agit donc de munir la machine d'une série de programmes, pour le calcul d'un choix assez riche de ce que nous appellerons *fonctions élémentaires ;* celles-ci seront suffisamment universelles, de façon que beaucoup d'autres fonctions puissent s'y ramener. C'est ainsi que le calcul de $\cos x$, $\operatorname{tg} x$ ... pourra se faire à partir de $\sin x$.

Le problème du choix des fonctions élémentaires pour une

(1) La machine ne tient aucun compte des décimales perdues, et ne « force » pas dans le cas où la partie coupée dépasse 5.

machine à virgule flottante mérite quelques commentaires. En premier lieu, il est bon que la fonction soit calculée avec une erreur relative de l'ordre de 10^{-14}, puisque la machine est capable de noter le résultat avec cette précision. Un exemple de calcul inadéquat est celui de cos x basé sur la série de Maclaurin. Pour x proche de $\pi/2$, la fonction est aussi petite que l'on veut, tandis qu'une série limitée commençant par 1, et dont les premiers termes sont du même ordre de grandeur et notés avec 15 décimales, ne saurait fournir de chiffres corrects au-delà du rang 10^{-15} ; le résultat est donc entaché d'une erreur absolue de l'ordre de 10^{-15}, et l'erreur relative peut devenir énorme. La difficulté disparaît au contraire si on calcule cos x par sin $(\pi/2 - x)$, sin x étant lui-même calculé par une série de Maclaurin. La difficulté réapparaîtrait pour $x = \pi$, mais le calcul dans l'intervalle $(\pi/2, \pi)$ se ramène à celui dans $(0, \pi/2)$. La fonction sin x est donc préférable si l'on choisit un développement de Maclaurin, parce que celui-ci procède autour de $x = 0$ qui est un zéro de la fonction. En conclusion, si une fonction a un zéro, le développement doit être fait autour de ce point.

Parmi les fonctions usuelles, il en est telles que l'exponentielle $(10^x = e^{2{,}302\cdots x})$ qui ne semblent pas faire difficulté à ce point de vue. Dans l'intervalle (0 à 1) auquel on se ramène, 10^x ne varie que de 1 à 10 et l'erreur relative ne diffère de l'erreur absolue que par un facteur 10 tout au plus. Un développement à erreur absolue inférieure à 10^{-15} paraît donc satisfaisant. Et cependant lorsqu'au cours d'un problème surgit un calcul intermédiaire tel que $10^x - 1$ pour x petit, l'erreur relative sur le résultat est énorme. En fait, pour avoir $10^x - 1$ avec une erreur relative de 10^{-14}, si le calcul se fait en soustrayant l'unité de 10^x, il faudrait connaître 10^x avec bien plus de quinze décimales. Mais tel n'est pas le cas si l'on prend comme fonction élémentaire la combinaison $10^x - 1$ en bloc ; cette fonction s'annule pour $x = 0$, et un développement en série autour de ce point convient d'après le principe de l'alinéa précédent. Au contraire, le calcul de 10^x comme $(10^x - 1) + 1$ n'aggrave pas l'erreur pour x voisin de zéro, mais bien pour les valeurs négatives élevées de x. Là c'est 10^x lui-même qu'il convient de prendre comme fonction.

L'argumentation ci-dessus peut sembler spécieuse et attirer l'objection suivante : pour calculer avec une erreur relative inférieure à 10^{-14} l'expression $10^{0,2} - 1{,}5848$ (où la quantité soustraite reproduit les cinq premiers chiffres de $10^{0,2}$) il faut un développement autour de 1,5848 ; des exigences de ce genre demanderaient finalement des développements de toutes les fonctions aux environs de tous les points. La réponse est, évidemment, qu'il y a des points naturels remarquables, et le calcul précis de différences du type $10^x - 1$ est indispensable dans de nombreux cas. Un utilisateur de la machine, sachant qu'elle calcule la fonction exponentielle, espère normalement pouvoir calculer sh x par

$$\operatorname{sh} x = (e^x - e^{-x})/2$$

et trouverait cependant une énorme erreur relative aux petits arguments. Au contraire, la précision relative est bien de l'ordre de 10^{-14} par

$$\operatorname{sh} x = [(e^x - 1) - (e^{-x} - 1)]/2. \tag{2}$$

En bref, dans une machine à virgule flottante, l'erreur relative s'aggrave nécessairement après soustraction de deux nombres très voisins ; le tout est de choisir les programmes et les fonctions de base de façon telle qu'on puisse éviter de devoir faire intervenir des différences de ce genre dans les problèmes usuels.

Ces réflexions quelque peu triviales n'ont été faites ici que pour justifier la méthode employée pour le calcul de l'exponentielle et du logarithme : on choisit $10^x - 1$ comme fonction de base pour les petits arguments, et 10^x pour les arguments de module élevé. En fait, les arguments élevés sont réduits par des formules du type $10^{a+b} = 10^a 10^b$, et, pour b suffisamment petit, on calcule 10^b par $(10^b - 1) + 1$, de sorte que le seul développement en série est celui de $10^x - 1$. Nous verrons que celui-ci ne coûte pas plus cher (en nombre de termes) pour une erreur relative imposée, que celui de 10^x pour une erreur absolue imposée. Au prix d'une complication minime (addition ou soustraction de l'unité) nous avons ainsi rendu possible le calcul d'expressions telles que (2).

Une souplesse supplémentaire du même genre est utile dans le calcul du logarithme. Aux environs de $x = 1$, log x est très petit, de l'ordre de $x - 1$, mais il n'y a guère de sens à le calculer avec une précision supérieure à celle avec laquelle $x - 1$ est donné. Si c'est x qui est donné, avec 15 décimales, $x - 1$ n'est connu qu'avec une erreur absolue de l'ordre de 10^{-15}, et on n'en demande pas plus pour $\log x$, bien que l'erreur relative sur $\log x$ soit alors mauvaise. Mais, ici de nouveau, on peut donner $y = 1 - x$ lui-même, le noter avec une erreur relative de l'ordre de 10^{-14}, et exiger la même erreur relative sur $\log (1 + y)$. Un cas où cela est utile est le calcul de

$$\text{arc th } y = \frac{1}{2} \log \frac{1 + y}{1 - y} \tag{3}$$

qu'il est possible d'obtenir avec une erreur relative convenable même pour les petits arguments, si c'est la fonction $\log (1 + y)$ qui est programmée, et non si c'est la fonction $\log x$.

3. Approximation a erreur relative imposée.

Les fonctions transcendantes élémentaires sont de deux types :

a) celles dont le développement de Maclaurin est à convergence rapide (séries factorielles) : $\sin x$, e^x,

b) celles dont le développement de Mclaurin converge lentement (séries du type harmonique) : arc tg x, $\log (1 + x)$.

Nous discuterons les approximations de ces deux genres de fonctions par des méthodes différentes, et nous indiquerons comment les procédés connus d'approximation à erreur absolue imposée se modifient en procédés à erreur relative imposée.

3.1. *Amélioration d'un développement de Maclaurin*

Pour les séries du type (a), il est devenu usuel de gagner quelques termes d'un développement de Maclaurin par une substitution de polynômes de Tchebycheff dont nous rappelons le principe. Soit

$$f(x) = a_0 + a_1 x + \ldots + a_n x^n \tag{4}$$

un développement de Maclaurin dont on connaît l'erreur absolue maximum e_0 en module, dans l'intervalle $(-h,\ h)$ choisi. On sait que le polynôme de Tchebycheff de degré n

$$T_n(x/h) = \cos\ n\ \text{arc}\ \cos\ x/h \tag{5}$$

est inférieur à l'unité en module dans le même intervalle et a comme plus haut coefficient $h^{-n}2^{n-1}$. Soustrayant

$$a_n h^n 2^{1-n} T_n(x/h)$$

de (4), on annule le terme en x^n et on produit une erreur supplémentaire inférieure en module à

$$e_1 = a_n h^n / 2^{n-1}. \tag{6}$$

On a ainsi obtenu un polynôme d'approximation de degré $n - 1$ et d'erreur $e_0 + e_1$, qui peut être inférieure à celle qu'aurait le polynôme de Maclaurin de degré $n - 1$. Dans le cas de fonctions paires ou impaires, le polynôme de Tchebycheff ne détruit pas la parité, et le degré baisse de deux unités. Pour des fonctions à convergence rapide (type *a*), il suffit en général d'arrêter la série de Maclaurin à un degré tel que l'erreur soit négligeable par rapport à celle demandée (2 à 3 ordres de mieux) et réduire le degré de quelques unités en appliquant plusieurs fois successivement l'artifice ci-dessus, en ne s'arrêtant que lorsque la somme des erreurs supplémentaires du type (6) risque de dépasser la tolérance. Il est bien connu que, pour h pas trop grand, on obtient ainsi des polynômes qui ne sont pas loin de l'optimum (erreur absolue minimum dans l'intervalle pour un degré donné).

3.2. *Approximation à erreur relative par développement de* f(x) /x.

Le procédé ci-dessus se transpose de diverses façons lorsqu'on désire ne pas dépasser une erreur relative imposée. Le plus simple est d'appliquer l'amélioration de Tchebycheff au polynôme $p(x)$ approchant $f(x)/x$, puisqu'il s'agit dans ce cas de fonctions s'annulant en $x = 0$, donc avec $a_0 = 0$ dans (4). Soit e l'erreur absolue qui en résulte, donnant

$$\left| \frac{f(x)}{x} - p(x) \right| \leqslant e \tag{7}$$

On voit que $f(x)$ est approché par le polynôme $xp(x)$ avec une erreur relative

$$\epsilon = \left| \frac{f(x) - xp(x)}{f(x)} \right| \leqslant e \left| \frac{x}{f(x)} \right| \leqslant e\mathrm{M} \tag{8}$$

où M est la borne supérieure de $x/f(x)$ dans l'intervalle, p. ex. $\pi/2$ pour le cas de $\sin x$ dans l'intervalle $(-\pi/2, \pi/2)$. On trouvera un exemple détaillé de ce procédé au § 8.1.

3.3. *Polynômes trigonométriques.*

Pour des développements lentement convergents, l'amélioration de Tchebycheff est possible pour un très grand nombre de termes, réduisant donc une série de Maclaurin de degré n (de l'ordre de 100 p. ex.) à un polynôme de degré m (de l'ordre de 10). Le calcul est alors pénible, et il est préférable de déterminer directement un développement en polynômes de Tchebycheff limité au degré m dès le départ, sans passer par l'intermédiaire de la série de Maclaurin (1). On sait d'ailleurs qu'on obtient ainsi la limite, pour n infini, des $n - m$ réductions successives qu'on ferait selon le § 3.1. Pour l'intervalle $(-h, h)$ le développement désiré s'obtient en posant $x = h \cos\theta$ et en développant $f(h\cos\theta)$ en série de Fourier,

$$f(h\cos\theta) = \mathrm{A}_0 + \mathrm{A}_1 \cos\theta + \mathrm{A}_2 \cos 2\theta + \dots \tag{9}$$

qui donne

$$f(x) = \mathrm{A}_0 + \mathrm{A}_1\mathrm{T}_1(x/h) + \mathrm{A}_2\mathrm{T}_2(x/h) + \dots \tag{10}$$

où les polynômes de Tchebycheff sont toujours définis par (5), Arrêtant (10) au terme en T_n, ce qui fournit pour $f(x)$ un polynôme d'approximation de degré n, on commet une erreur absolue

$$e \leqslant |\mathrm{A}_{n+1}| + |\mathrm{A}_{n+2}| + \dots \tag{11}$$

(1) S. Bernstein, *Sur l'ordre de la meilleure approximation des fonctions continues par des polynômes de degré donné*, Mém. Acad. Roy. Belge, Cl. des Sc., 2e série, t. IV, 1922 ; C. Lanczos, *Trigonometric Interpolation of Empirical and Analytical Functions*, Journ. Math. Phys., vol. XVII, 1938, p. 123.

3.4. *Polynômes trigonométriques à erreur relative : premier procédé.*

Par le procédé du § 3.2. un développement à erreur relative de $f(x)$ se déduit d'un développement à erreur absolue de $f(x)/x$, et la même modification vaut pour (10). On remarque que le développement en série de Fourier de $f(x)/x$ se déduit de celui de la dérivée $f'(x)$ par intégration par rapport à h. En effet

$$\int_0^h f'(h\cos\theta)dh = \frac{1}{\cos\theta}\int_0^h f'(h\cos\theta)d(h\cos\theta) =$$

$$\frac{f(h\cos\theta)}{\cos\theta} = h\frac{f(x)}{x}. \tag{12}$$

Ainsi le développement à erreur relative de $\log(1+x)$, qui demanderait celui en série de Fourier de

$$\frac{\log(1+h\cos\theta)}{h\cos\theta}$$

s'obtient en fait par l'intégration de la série de Fourier de $(1+h\cos\theta)^{-1}$. Les coefficients de cette série sont cependant des fonctions compliquées de h ; bien que ces fonctions soient intégrables explicitement, les expressions finales des coefficients sont peu propices au calcul numérique.

3.5. *Second procédé.*

Un autre procédé, bien que légèrement moins convergent, est alors plus commode. Partant d'une série (10) pour $f'(x)$ avec une erreur $e(x)$ inférieure à e en module, qui conduit après regroupement des termes à un polynôme de degré n du type (4), on intègre les deux membres par rapport à x ; il vient

$$f(x) = a_0x + \frac{a_1x^2}{2} + \ldots + \frac{a_nx^{n+1}}{(n+1)} + \int_0^x e(x)dx. \tag{13}$$

Mais

$$\left|\int_0^x e(x)dx\right| < \int_0^x |e|dx = |e|x. \tag{14}$$

Appelant $p(x)$ le polynôme au second membre de (13) divisé par x (donc de nouveau de degré n), on retrouve (7), donc (8). On obtient ainsi encore une fois un polynôme d'approximation $p(x)$ pour $f(x)/x$ à partir d'un développement en série de Fourier de $f'(h \cos \theta)$; il n'y a cependant plus d'intégration en h, mais bien en x, et celle-ci est élémentaire.

En résumé, nous avons établi deux procédés pour passer d'un polynôme d'approximation en valeur absolue à un polynôme d'approximation en erreur relative. On voit facilement qu'en partant d'un développement de Maclaurin non amélioré, les deux procédés (§ 3.2. et § 3.5) ne sont pas distincts. Ils le deviennent dès que les polynômes de Tchebycheff interviennent, et la discussion de quelques exemples simples montre que le second procédé est en général un peu moins bon ; comme la différence est cependant minime, et qu'il offre l'avantage d'un calcul facile des coefficients, il a été préféré.

4. Quelques développements en série de Fourier.

Le calcul de $\log (1 + x)$ et de $\operatorname{arc\,tg} x$ par le procédé des § 3.4. et 3.5. demande le développement en polynômes de Tchebycheff des dérivées $(1 + x)^{-1}$ et $(1 + x^2)^{-1}$. Posant $z = re^{i\theta}$ dans

$$\frac{1}{1+z} = 1 - z + z^2 - z^3 + \ldots \tag{15}$$

et prenant les parties réelles, il vient

$$\frac{1 + r \cos \theta}{1 + r^2 + 2r \cos \theta} = 1 - r \cos \theta + r^2 \cos 2\theta \ldots$$

soustrayant $\frac{1}{2}$ des deux membres, et multipliant par $2(1 + r^2)/(1 - r^2)$ on a le développement désiré de $(1 + x)^{-1}$:

$$\frac{1}{1 + h \cos \theta} = \frac{1 + r^2}{1 - r^2}(1 - 2r \cos \theta + 2r^2 \cos 2\theta \ldots) \tag{16}$$

avec

$$h = \frac{2r}{1 + r^2} \tag{17}$$

Pour la meilleure convergence de (16) on choisira r comme la plus petite solution de l'équation du second degré (17), donc

$$r = \frac{1 - \sqrt{1 - h^2}}{h} \tag{18}$$

L'erreur (11) est, si (16) est arrêté au terme en r^n,

$$e = \frac{1 + r^2}{1 - r^2} (2r^{n+1} + 2r^{n+2} + \dots) = \frac{2(1 + r^2)r^{n+1}}{(1 - r^2)(1 - r)} \tag{19}$$

Par

$$\frac{1}{1 + h \cos \theta} + \frac{1}{1 - h \cos \theta} = \frac{2}{1 - h^2 \cos^2 \theta}$$

et changement de h en ih, et changement corrélatif de r en ir, on obtient le développement de $(1 + x^2)^{-1}$, à savoir

$$\frac{1}{1 + h^2 \cos^2\theta} = \frac{1 - r^2}{1 + r^2} (1 - 2r^2 \cos 2\theta + 2r^4 \cos 4\theta \dots) \tag{20}$$

avec

$$r = \frac{\sqrt{1 + h^2} - 1}{h} \tag{21}$$

l'erreur (11) étant alors

$$e = \frac{1 - r^2}{1 + r^2} (2r^{n+2} + 2r^{n+4} + \dots) = \frac{2r^{n+2}}{1 + r^2} \tag{22}$$

lorsque (20) est arrêté au terme en r^n.

5. Erreurs d'arrondi dans le calcul d'un polynôme.

5.1. *Arrondis dans les équations élémentaires.*

Dans une machine à virgule flottante à mantisse de 15 chiffres, l'erreur d'arrondi due à la multiplication est au plus une unité par défaut au dernier rang, que l'arrondi provienne de ce qu'on a coupé la queue du produit pour le ramener à quinze chiffres, ou du réalignement pour faire rentrer dans la mantisse le report de tête, ou des deux causes à la fois. Ceci donne, dans le cas le plus défavorable, une erreur relative $\epsilon = 10^{-14}$. Cette erreur est toujours par défaut sur le module du résultat.

Dans l'addition ou la soustraction, il n'y a pas d'erreur si les termes ont même exposant et qu'il n'y a pas de réalignement (report ou apparition d'un zéro en tête) S'il y a report en tête, le dernier chiffre du résultat est perdu, ce qui donne de nouveau une erreur par défaut. Si les termes n'ont pas le même exposant et s'il ne s'agit pas de la soustraction de deux nombres voisins, celui qui a le plus petit exposant est aligné sur l'autre avant l'opération, et sa queue est perdue ; ceci donne encore la même erreur relative, mais cette fois par excès ou par défaut. Même si les deux causes agissent simultanément (p. ex. dans $0{,}999 \ldots + 0{,}1 \,.\, 10^{-15}$) l'erreur relative maximum due à l'arrondi ne dépase pas ϵ en module.

Dans la soustraction de deux nombres voisins ayant même exposant, il n'y a pas d'erreur dans le résultat si les données sont exactes. C'est par le fait que les données sont elles-mêmes affectées d'erreurs qu'un effet héréditaire pouvant produire une erreur relative considérable se transmet à la différence. L'erreur de l'opération proprement dite est cependant nulle.

Dans la soustraction de deux nombres voisins dont les exposants sont différents (au plus d'une unité puisque les nombres sont voisins, p. ex. $0{,}1 \,.\, 10^1 - 0{,}9 \,.\, 10^0$) il y a alignement du plus petit sur le plus grand avant soustraction, donc une erreur absolue pouvant atteindre 10^{p-15}, où p est l'exposant du nombre qui sera altéré par alignement. Comme la différence peut être très petite (p. ex. dans $0{,}100 \ldots 10^1 - 0{,}999 \ldots 10^0$) l'erreur relative peut devenir énorme. Dans tous les cas, cette erreur a toutefois son origine dans l'erreur absolue 10^{p-15} sur celui des deux termes qui est le plus petit en module, et ceci correspond à une erreur relative ϵ par défaut sur ce terme. Dans le calcul d'un polynôme, où les multiplications alternent avec les additions ou soustractions, les termes sont déjà affectés d'erreurs du même ordre, et nous allons montrer que ces erreurs ne cumulent jamais.

Il suffit pour cela de considérer l'opération binôme $ax - b$. On a vu que l'erreur relative sur ax était au plus ϵ par défaut. Si $|b| < |ax|$, l'alignement de b produit une erreur du même ordre et dans le même sens, et les erreurs ont donc tendance à se compenser ; on obtient donc certainement une majoration en comptant ϵ par opération. Si $|ax| \leqslant |b|$, c'est ax qui est aligné et dont la queue est amputée ; mais ax a déjà subi une mutilation analogue après la multiplication qui l'a constitué, et en majorant à ϵ le module de l'erreur relative qui en résultait, on a déjà tenu compte de l'amputation maximum.

En conclusion, on peut calculer comme si toute opération élémentaire donnait une erreur relative ne dépassant pas ϵ en module, donc multipliait le résultat par $\eta = 1 \pm \epsilon$ tout au plus, à condition qu'il n'y ait jamais une succession immédiate de deux soustractions.

5.2. *Calcul d'un polynôme.*

Le calcul de $f(x) = ax^4 + bx^3 + cx^2 + dx + e$ se fait dans l'ordre

$$\{[(ax + b)x + c]x + d\}x + e.$$

Des facteurs d'erreur η_i s'introduisent successivement dans les résultats partiels

$$ax\eta_1 \; ; \; (ax\eta_1 + b)\eta_2 \; ; \; [(ax\eta_1 + b)\eta_2]x\eta_3 \ldots$$

où nous avons distingué les divers η car leurs signes varient. Finalement le terme en x^i du polynôme est multiplié par H_i où

$$H_0 = \eta_8 \; ; \; H_1 = \eta_6\eta_7H_0 \; ; \; H_2 = \eta_4\eta_5H_1 \; ;$$
$$H_3 = \eta_2\eta_3H_2 \; ; \; H_4 = \eta_1H_3.$$

Chaque élévation de degré comporte deux opérations, de sorte que le terme en x^i est multiplié par $2i + 1$ facteurs η, sauf le terme de plus haut degré qui a un facteur η en moins, car il n'y a pas d'erreur initiale. Bien que les valeurs des coefficients H_i ne soient pas complètement indépendantes, la seule façon simple d'obtenir une majorante de l'erreur est de considérer la combinaison la plus défavorable où toutes les erreurs cumulent dans le même sens.

Si x et tous les coefficients sont positifs, la majorante s'obtient en posant $\eta = 1 + \epsilon$ partout et le polynôme devient

$$\eta[(a/\eta)(x\eta^2)^4 + b(x\eta^2)^3 + c(x\eta^2)^2 + d(x\eta^2) + e].$$

Remplaçant a/η par a, ce qui ne fait que majorer l'erreur, cette expression devient

$$\eta f(x\eta^2) = (1 + \epsilon)f(x + 2\epsilon x) = f(x) + \epsilon f(x) + 2\epsilon x f'(x),$$

donnant une erreur relative en module

$$(1 + 2xf'/f)10^{-14}. \tag{23}$$

Si x ou certains coefficients sont négatifs, il suffit de prendre les valeurs absolues de tous les coefficients et de x dans la correction $\epsilon f + 2\epsilon x f'$. Appelant $F(x)$ le polynôme dont les coefficients sont les valeurs absolues de ceux de $f(x)$, posant $X = |x|$, et écrivant F sans plus pour $F(X)$, la correction est $\epsilon F + 2\epsilon X F'$, et l'erreur relative devient

$$10^{-14}(F + 2XF')/f. \tag{24}$$

5.3. *Erreurs dues à l'arrondi sur les coefficients.*

Une erreur supplémentaire provient de ce que les coefficients du développement en série sont eux-mêmes arrondis à 15 chiffres significatifs dans la mémoire de la machine. Cela provoque sur chaque terme une erreur relative dont le module est au maximum $0{,}5.10^{-14}$, si l'arrondi a été fait en forçant avant la mise en mémoire. Si tous les termes sont de même signe, on obtient une majorante en admettant que chaque terme est multiplié par le même facteur, ce qui multiplie le résultat par le même facteur. L'erreur relative sur le résultat est alors $0{,}5 \,.\, 10^{-14}$.

Si les signes ne sont pas constants, le polynôme f est à remplacer par F comme ci-dessus dans l'erreur absolue, et l'erreur relative est

$$0{,}5 \,.\, 10^{-14}\, F/f. \tag{25}$$

5.4. *Erreur totale.*

Par combinaison de (23) ou (24) et de l'erreur d'arrondi des coefficients on trouve, posant $\epsilon = 10^{-14}$,

$$\epsilon(1{,}5 + 2xf'/f)$$

pour le cas d'une série positive, et

$$\epsilon(1{,}5 \,.\, F + 2XF')/f \tag{26}$$

pour le cas général.

Dans le cas d'une série sans terme constant, la formule (26) est à appliquer à $g = f/x$, la multiplication ultérieure par x pour passer à $f(x)$ donnant une erreur relative ϵ supplémentaire, à combiner en module. Posant $G = F(X)/X$, l'erreur totale devient

$$\epsilon[(1{,}5\ G + 2XG')/|g| + 1] = \epsilon[(-0{,}5F + 2XF')/|f| + 1] \quad (27)$$

Pour une série paire, la variable est $y = x^2$, et comporte en elle-même une erreur relative ϵ due à l'élévation au carré. Cette erreur ϵ sur y donne une erreur héréditaire sur $f(x) = \varphi(y)$, valant, compte tenu de $dy = 2x\ dx$, $\epsilon x f'/2f$. L'erreur proprement dite d'arrondi (26) sur φ s'y ajoute. Écrivant Y et Φ avec la convention usuelle, cette erreur est

$$\epsilon(1{,}5\,\Phi + 2Y\Phi')/\varphi = \epsilon(1{,}5F + XF')/f$$

ce qui donne une erreur totale

$$\epsilon(1{,}5F + XF' + 0{,}5X|f'|)/f. \quad (28)$$

Pour une série impaire, le passage de (26) à (27) est en outre à faire sur (28), ce qui donne

$$\epsilon[(0{,}5F + XF' + 0{,}5|f - xf'|)/|f| + 1]. \quad (29)$$

5.5. *Applications.*

Pour les fonctions telles que $\log(1 + x)$, $e^x - 1$, à développement dans un intervalle assez petit aux environs de $x = 0$, on a $f \mathrel{\hat{=}} x$ et la formule (27) donne une erreur relative de $2{,}5\ \epsilon$.

Pour les fonctions telles que arc tg x, la même simplification dans (29) donne aussi $2{,}5\ \epsilon$.

Pour $\sin x$, l'intervalle de base $(-\pi/2, \pi/2)$ est trop grand pour une telle approximation. Posant $F(X) = \operatorname{sh} X$ dans (29), l'erreur est

$$\epsilon\left[\frac{0{,}5\ \operatorname{sh} X + X \operatorname{ch} X}{|\sin x|} + 0{,}5|1 - x \operatorname{cotg} x| + 1\right].$$

Chacun des termes est maximum pour $x = \pi/2$ et on a finalement

$$\epsilon(0{,}5\ \operatorname{sh} \pi/2 + \pi/2\ \operatorname{ch} \pi/2 + 1{,}5) = 6{,}58\ \epsilon. \quad (30)$$

6. Calcul de l'inverse.

Nous avons vu que tout le problème était ramené au calcul de $y = 1/x$ dans l'intervalle de 0,1 à 1 pour x. Le calcul se fait à partir d'une première approximation y_0 donnée par un polynôme et les améliorations données par l'itération

$$y_n = y_{n-1}(a_n - xy_{n-1}) \tag{31}$$

où a_n est la constante 2 dans l'itération de Newton-Raphson, et une valeur légèrement supérieure dans la modification introduite par F. Storrer [1].

Lorsque le degré du polynôme d'approximation y_0 croît, le nombre d'itérations requis pour arriver à la précision relative de 10^{-14} décroît, et, pour un certain degré, le nombre d'instructions élémentaires composant le programme total est minimum. Quelques tâtonnements préliminaires ont indiqué que ce minimum se produisait pour un polynôme du premier degré, et qu'il fallait ensuite cinq itérations. L'amélioration de Storrer n'a été appliquée qu'aux trois premières, car, même ainsi, l'erreur reste en dessous de 10^{-14} après la dernière.

On sait [1] que pour l'intervalle (a, b) l'approximation linéaire optimum en valeur relative est

$$\frac{1}{x}\left[1 - \frac{T_2\left(\dfrac{a+b-2x}{b-a}\right)}{T_2\left(\dfrac{b+a}{b-a}\right)}\right] = \frac{8(a+b-x)}{a^2+6ab+b^2} \tag{32}$$

et l'erreur relative maximum qui en résulte est

$$e_0 = \frac{1}{T_2\left(\dfrac{a+b}{b-a}\right)} = \frac{(b-a)^2}{a^2+6ab+b^2}. \tag{33}$$

A partir de e_{n-1}, (n quelconque), une formule de Storrer détermine le coefficient a_n à adopter au stade suivant dans (31). Ce coeffi-

(1) F. Storrer, *Amélioration du procédé de division utilisant l'itération de Newton-Raphson*, Bull. Acad. Roy. Belg., Classe des Sc., 5e série, t. XLII, pp. 30-33, 1956.

cient n'a été calculé qu'avec la précision déjà nécessaire au stade correspondant, et il en résulte une certaine erreur relative e_n après la n-ième itération. Cette erreur a été calculée en tenant compte de ce que a_n n'est qu'approximativement le a_n optimum. Calculant e_0 par (33) avec $a = 0{,}1$ et $b = 1$, on trouve $e_0 = 0{,}50310559$; $e_1 = 0{,}1713412$; $e_3 = 0{,}0160836$; $e_3 = 0{,}0001304$.

Les deux dernières itérations se font avec $a_4 = a_5 = 2$, et les erreurs correspondantes se déduisent de la règle $e_{n+1} = e_n^2$. On trouve

$$e_4 = 1{,}7 \,.\, 10^{-8} \,;\; e^5 = 3 \,.\, 10^{-16}.$$

Les constantes numériques nécessaires au calcul sont rassemblées à la Table 1, où a_0 et b_0 représentent les coefficients du polynôme (32) noté $a_0 x + b_0$, les constantes $a_1 a_2 a_3$ étant celles des trois premières itérations (31).

Le contrôle du calcul de l'inverse $y = 1/x$ se fait en vérifiant que $xy - 1$ est suffisamment petit. L'erreur mathématique maximum se produit pour $x = 0{,}1$ donc $y = 10$ et résulte de l'erreur relative e_5 évaluée ci-dessus ; il s'ensuit $|xy - 1| \leqslant 3.10^{-16}$ Une analyse détaillée montre que cette erreur ne se cumule pas avec celles, plus importantes, dues à l'arrondi. Sur y lui-même, on peut établir que l'erreur relative due à l'arrondi ne dépasse pas 10^{-14} en module. Dans le calcul de $xy - 1$ il peut apparaître un arrondi semblable, de sorte que la formule de contrôle à adopter est

$$|xy - 1| \leqslant 2 \cdot 10^{-14}. \tag{34}$$

7. Calcul de la racine carrée.

On sait que le procédé de Newton-Raphson pour $\sqrt{x}$ conduit à une formule du type (31) impliquant une division. Par contre, le procédé appliqué à $1/\sqrt{x}$ donne

$$y_n = y_{n-1}(a_n - 0{,}5 x y_{n-1}^2) \tag{35}$$

avec $a_n = 1{,}5$ (nous trouverons une valeur légèrement supérieure selon le procédé amélioré), et il est donc commode d'adopter $y = 1/\sqrt{x}$ comme fonction de base, quitte à calculer $\sqrt{x}$ par xy.

Nous supposons toujours x compris entre 0,1 et 1, car pour l'intervalle de 1 à 10 on s'y ramène par multiplication du résultat par $\sqrt{10}/10$, et pour les puissances paires de 10 on multiplie simplement l'exposant par — 0,5. On obtient un programme à nombre minimum d'instructions à partir d'un polynôme d'approximation en erreur relative du second degré $y_0 = ax^2 + bx + c$ dont les coefficients (donnés dans la table II) ont été déterminés par tâtonnements : les écarts relatifs entre le polynôme et $1/\sqrt{x}$ oscillent entre les extrêmes suivants :

$$-0{,}097\ ; \quad +0{,}097465\ ; \quad -0{,}094278\ ; \quad +0{,}097.$$

On a donc

$$e_0 < 0{,}0975. \tag{36}$$

Dans le cas où une multiplication par $\sqrt{10}$ est requise, il est avantageux de l'effectuer avant les itérations : ceci dispense de connaître $\sqrt{10}$ avec précision, puisque le résultat sera automatiquement amélioré par itérations. Quoi qu'il en soit, la valeur précise de $\sqrt{10}$ est donnée dans la table II.

L'amélioration de (35) par $a_n = 1{,}5 + \delta_n$ a pour but de symétriser l'erreur relative, de telle sorte que, de $-e_n < \epsilon_{n-1} < e_{n-1}$, on puisse déduire $-e_n < \epsilon_n < e_n$, ϵ désignant toujours l'erreur relative réelle et e la borne de son module. Un calcul élémentaire montre que

$$\epsilon_n = \delta_n(1 + \epsilon_{n-1}) - \frac{3}{2}\,\epsilon_{n-1}^2 - \frac{1}{2}\,\epsilon_{n-1}^3 \tag{37}$$

et on détermine δ_n de façon que (37) soit la meilleure approximation de zéro au sens de Tchebycheff dans l'intervalle $(-e_{n-1}, e_{n-1})$ de la variable ϵ_{n-1}. Une discussion assez longue sur les écarts extrêmes, semblable à celle de la référence, montre que la valeur optimum pour δ_n, et la borne e_n de $|\epsilon_n|$ qui en résulte, sont données par

$$\delta_n = \frac{3}{4}\,e_{n-1}^2 + \frac{1}{8}\,e_{n-1}^3 + \frac{1}{64}\,e_{n-1}^4 - \frac{1}{128}\,e_{n-1}^5 - \frac{5}{1536}\,e_{n-1}^6 + \dots \tag{38}$$

$$e_n = \frac{3}{4}\,e_{n-1}^2 + \frac{1}{8}\,e_{n-1}^3 + \frac{7}{64}\,e_{n-1}^4 + \frac{3}{128}\,e_{n-1}^5 - \frac{7}{1536}\,e_{n-1}^6 + \dots \tag{39}$$

Nous n'avons employé cette amélioration que pour la première itération : (36) et (39) donnent alors

$$e_1 < 0{,}007256. \qquad (40)$$

On continue par trois itérations ordinaires. On calcule e_2 par (37), avec $\delta_n = 0$ et la valeur (40) de $e_1 = \max |\epsilon_1|$. Pour les itérations suivantes, on sait que ϵ_{n-1} est négatif de sorte que les deux derniers termes de (37) sont nécessairement de signes opposés. On calcule donc e_3 et e_4 par

$$e_n = (-3e_{n-1}^2 + e_{n-1}^3)/2.$$

Ceci donne

$$e_2 = 0{,}7917 \cdot 10^{-4}\,; \quad e_3 = 0{,}9401 \cdot 10^{-8}\,; \quad e_4 = 0{,}133 \cdot 10^{-15}.$$

Les constantes sont mentionnées au tableau II.

Une analyse détaillée des erreurs d'arrondi montre que l'erreur relative finale e sur y est donnée par

$$-1{,}0133 \cdot 10^{-14} < e < 1{,}75 \cdot 10^{-14} \qquad (41)$$

et l'on démontre que la tolérance à adopter sur le contrôle est

$$|xy^2 - 1| \leq 4.10^{-14}. \qquad (42)$$

8. Calcul de sin x.

8.1. *Développement en série.*

Nous supposons l'argument préalablement réduit dans l'intervalle $(-\pi/2, \pi/2)$. La série de Maclaurin de $(\sin x)/x$ arrêtée au terme x^{18}

$$f_1(x) = 1 - \frac{x^2}{3!} + \frac{x^4}{5!} - \ldots - \frac{x^{14}}{15!} + \frac{x^{16}}{17!} - \frac{x^{18}}{19!}$$

y représente $(\sin x)/x$ avec une erreur absolue inférieure en module à

$$e_1 = (\pi/2)20/21! = 1{,}6\,.\,10^{-16}.$$

Pour annuler le terme en x^{18} on ajoute

$$f_2(x) = \frac{T_{18}(2x/\pi)}{19!\,2^{17}}\left(\frac{\pi}{2}\right)^{18} = \frac{1}{19!}\left[x^{18} - \frac{9}{2}\left(\frac{\pi}{2}\right)^2 x^{16} + \frac{135}{16}\left(\frac{\pi}{2}\right)^4 x^{14} \ldots\right]$$

qui ajoute une erreur

$$e_2 = (\pi/2)^{18}/2^{17}19! = 8{,}2\,.\,10^{-21}. \tag{43}$$

Le résultat est un polynôme terminé en x^{16} et de plus haut coefficient

$$C = \frac{1}{17!} - \frac{9}{2}\left(\frac{\pi}{2}\right)^2 \frac{1}{19!} = \frac{1}{17!}\left[1 - \frac{1}{76}\left(\frac{\pi}{2}\right)^2\right] = 2{,}7\,.\,10^{-15} \tag{44}$$

Soustrayant encore un polynôme de Tchebycheff de coefficient C pour annuler le terme en x^{16}, c'est-à-dire

$$f_3(x) = \frac{CT_{16}(2x/\pi)}{2^{15}}\left(\frac{\pi}{2}\right)^{16} = C\left[x^{16} - 4\left(\frac{\pi}{2}\right)^2 x^{14} + \frac{13}{2}\left(\frac{\pi}{2}\right)^4 x^{12} - \ldots\right]$$

ce qui produit une erreur

$$e_3 = C(\pi/2)^{16}/2^{15} = 1{,}1\,.\,10^{-16} \tag{45}$$

on obtient un polynôme.

$$f_4(x) = f_1(x) + f_2(x) - f_3(x) = a_1 + a_3x^2 + \ldots\, a_{15}x^{14}$$

approchant $(\sin x)/x$ dans $(-\pi/2, \pi/2)$ avec une erreur absolue inférieure à

$$e = e_1 + e_2 + e_3 = 2{,}7\,.\,10^{-16}. \tag{46}$$

Si l'on adopte pour le calcul de $\sin x$ le polynôme

$$x\, f_4(x) = a_1x + a_3x^3 + \ldots + a_{15}x^{15} \tag{47}$$

l'erreur relative sera (8), soit

$$\epsilon = \pi e/2 = 4{,}3\,.\,10^{-16}. \tag{48}$$

8.2. *Contrôle.*

Le contrôle se fait en calculant $z = \sin x/3$ et en vérifiant que l'expression $y - z(3 - 4z^2)$ s'annule [1]. Le calcul n'est précis en

[1] Avec $z = \sin x/2$, la formule de contrôle, rendue rationnelle, serait $y^2 - 4z^2(1 - z^2) = 0$, demanderait plus d'opérations et mènerait à une tolérance supérieure à (50).

erreur relative que si z s'annule lorsque y s'annule [1] ; il en sera certainement ainsi si l'on calcule $x/3$ après réduction de x à l'intervalle $(-\pi/2,\ \pi/2)$.

Pour des erreurs relatives ϵ sur y et η sur z, on obtient

$$y - z(3 - 4z^2) = z[3(\epsilon - \eta) + 4(3\eta - \epsilon)z^2]. \tag{49}$$

Par (30), on a $|\epsilon| < 6{,}58 \,.\, 10^{-14}$. Mais les arrondis dans le calcul de $z(3 - 4z^2)$ donnent, tous calculs faits, une erreur relative dont le module peut atteindre 10^{-14} (compte tenu de ce que $3 - 4z^2$ vaut au moins 2 pour $|z| \leqslant \sin \pi/6$), ce qui équivaut à ajouter $z(3 - 4z^2)10^{-14}$ au second membre de (49), donc porter ϵ à $7{,}58 \,.\, 10^{-14}$. D'autre part la multiplication par $1/3$ donne une erreur de 10^{-14} sur $x/3$, donc une erreur héréditaire $10^{-14}(x/3)$ cotg $x/3$, valant pratiquement 10^{-14}, sur sin $x/3$. Cette erreur s'ajoute à celle $6{,}58 \,.\, 10^{-14}$ du calcul de z pour donner $|\eta| < 7{,}58 \,.\, 10^{-14}$. Pour ces valeurs la combinaison des signes la plus défavorable dans (49), pour l'intervalle $|z| < \sin \pi/6$, est ϵ positif et η négatif ; et la borne supérieure du module du crochet est atteinte pour $z = 0$ et vaut $4{,}55 \,.\, 10^{-13}$. D'où la formule de contrôle

$$|y - z(3 - 4z^2)| < 4{,}55 \,.\, 10^{-13}|z| \tag{50}$$

Remarquons enfin que, pour $x < 10^{-25}$, le calcul de la série (47) devient impossible, car la variable x^2 qui y intervient est inférieure à 10^{-50} et déclenche l'alarme d'exposant. En fait pour $x < 10^{-8}$, on a $\sin x = x$ avec une erreur relative inférieure à 10^{-15}, et c'est cette formule qu'il convient d'utiliser. Les valeurs proches d'un multiple de π ne font alors pas de difficulté, car on ne peut noter avec 15 décimales que les valeurs qui diffèrent d'un tel multiple de 10^{-14} au minimum. On a alors $x > 10^{-14}$ après réduction au premier quadrant.

8.3. *Réduction à l'intervalle de base.*

Examinons de plus près la façon dont se fait la réduction de x à l'intervalle $(-\pi/2,\ \pi/2)$. Partant de x, on commence par écrire

[1] Pour cette raison on ne peut faire le contrôle à partir de sin $3x$.

$x_1 = x/2\pi$ et à en garder la partie fractionnaire, appelée x_2, qui conserve le signe de x_1. On a ainsi $|x_2| \leqslant 1$ et x_2 représente l'argument en fraction de circonférence ; de plus $\sin x = \sin 2\pi x_2$. Pour $|x_2| \leqslant 0{,}25$, on est déjà dans l'intervalle $(-\pi/2, \pi/2)$. Pour $|x_2| > 0{,}25$, la réduction à $|x_3| < 0{,}25$ se fait par

$$x_3 = [||x_2| - 0{,}75| - 0{,}25] \operatorname{sgn} x_2 \qquad (51)$$

et on a $\sin x = \sin 2\pi x_3$. La formule (46) se justifie comme suit :

(*a*) pour $|x_2| < 0{,}75$, elle donne $x_3 = 0{,}5 - x_2$ pour $x_2 > 0$ et $x_3 = -0{,}5 - x_2$ pour $x_2 < 0$; dans les deux cas elle ramène l'argument des quadrants 2 ou 3 vers les quadrants 1 ou 4 par prise du supplément.

(*b*) pour $|x_2| > 0{,}75$, (51) donne $x_3 = x_2 - 1$ pour $x_2 > 0$ et $x_3 = x_2 + 1$ pour $x_2 < 0$; elle ramène donc l'argument du quatrième au premier quadrant.

Finalement, il y a lieu de calculer $\sin 2\pi y$, avec $y = x_2$ pour $|x_2| \leqslant 0{,}25$, et $y = x_3$ pour $|x_2| > 0{,}25$. La série (47) est donc à utiliser avec l'argument $2\pi y$ et devient

$$\sin 2\pi y = b_1 y + b_3 y^3 + \dots + b_{15} y^{15} \qquad (52)$$

où

$$b_i = (2\pi)^i a_i.$$

Ce sont ces coefficients b_i, rebaptisés a_i, qui sont donnés à la table III. Remarquons enfin qu'il aurait été possible d'englober également le cas $|x| \leqslant 0{,}25$ dans une formule compacte du type (51) au prix d'une légère complication. Ceci serait techniquement avantageux car l'exécution de (51) ne demande, outre les opérations élémentaires, que les altérations « module » et « signature », ne prenant pas de temps. Au contraire, la séparation initiale du cas $|x| \leqslant 0{,}25$, et le by-pass des réductions ultérieures intervenant dans le cas opposé, demandent un saut conditionnel dans le programme. La technique du saut a quand même été adoptée pour la première alternative parce qu'une formule du type (51) forcerait à calculer intermédiairement x au moyen d'une écriture équivalant à $\pi/2 - (\pi/2 - x)$ pour les petits arguments, ce qui détruirait la précision.

9. Calcul de arc tg x.

9.1. *Réduction et développement en série.*

Pour x négatif, on se ramène à x positif par changement de signe ; pour $x > 1$, on passe à l'inverse, qui est inférieur à l'unité, par

$$\text{arc tg}\, x = \pi/2 - \text{arc tg}\, 1/x. \tag{53}$$

Le calcul de $y = \text{arc tg}\, x$ dans $0 < x < 1$ nécessiterait un polynôme trop long. On le limite à $0 < x < h$ en déterminant h de façon que la substitution

$$\text{arc tg}\, x = \text{arc tg}\, w + \text{arc tg}\, a \tag{54}$$

permette le calcul de arc tg w, où

$$w = \frac{x - a}{1 + xa} \tag{55}$$

par le même développement en série. Puisque celui dans l'intervalle $(0, h)$ est impair, il convient en fait dans $(-h, h)$ et c'est dans cet intervalle de w que l'intervalle $(h, 1)$ de x doit être transformé par (55). Ceci donne deux équations fixant a et h. On trouve que h est solution de $1 - 4h + h^2 = 0$, où il est avantageux de prendre la plus petite racine

$$h = 2 - \sqrt{3} = 0{,}2679 \ldots \tag{56}$$

et on obtient

$$a = \frac{1 - h}{1 + h} = \frac{1}{\sqrt{3}} = 0{,}5773 \ldots \tag{57}$$

La valeur intervenant dans (57) est

$$\text{arc tg}\, a = \text{arc tg}\, 1/\sqrt{3} = \pi/6 \tag{58}$$

Pour le développement dans $(-h, h)$ on adopte le procédé du § 3.5, à partir de (20). La valeur de r^2 résultant de (21) est 0,01733... et l'erreur (22) vaut

$$e = 2{,}6 \,.\, 10^{-16} \tag{59}$$

pour $n = 16$. L'expression (20) donne donc un polynôme pair en x du 16-*ième* degré, et après intégration en x, on obtient le polynôme d'approximation cherché

$$\text{arc tg}\, x = a_1 x + a_3 x^3 + \ldots + a_{17} x^{17} \tag{60}$$

dont l'erreur relative est toujours (59), grâce à (8), où $M = 1$.

9.2. *Erreurs et contrôle.*

(*a*) Dans l'intervalle de base $|x| < h$, l'erreur est celle d'arrondi calculée au § 5.5., donc approximativement $2{,}5 \,.\, 10^{-14}$. A un degré d'approximation au delà, (29) donne $\epsilon(2{,}5 + 2h^2)$, soit $2{,}65 \,.\, 10^{-14}$.

(*b*) Pour $1 > x > h$, il apparaît des erreurs supplémentaires dues à la réduction à l'intervalle de base. Dans le calcul de $1 + xa$, l'arrondi ne dépasse pas ϵ. L'inversion ajoute ϵ et l'erreur relative sur $(1 + xa)^{-1}$ est donc 2ϵ. L'erreur absolue sur $x - a$ est uniquement due à l'arrondi sur a, soit $0{,}05\epsilon$, et l'erreur absolue sur w s'obtient par

$$\begin{aligned} \Delta w &= (x - a)\Delta(1 + xa)^{-1} + (1 + xa)^{-1}\Delta(x - a) \\ &= \frac{2|x - a| + 0{,}05}{|1 + xa|}. \end{aligned}$$

Comme $x < 1$, on a $|x - a| \leqslant 0{,}43$; d'autre part $|1 + xa| > 1 + ha = 1{,}15$, de sorte que $\Delta w < 0{,}8\epsilon$. A cela s'ajoute ϵ en erreur relative pour l'arrondi de la multiplication de $x - a$ par $(1 + xa)^{-1}$; comme $w < h$, l'erreur absolue correspondante est au plus $0{,}1\epsilon$, et l'erreur absolue totale sur w ne dépasse pas $0{,}9 \,.\, 10^{-14}$. L'erreur absolue que ceci induit sur arc tg w est multipliée par $(1 + w^2)^{-1}$, donc inférieure. A cela s'ajoute l'erreur relative $2{,}65 \,.\, 10^{-14}$ de calcul de arc tg w, correspondant au maximum à une erreur absolue de $2{,}65 \,.\, 10^{-14}$ arc tg $h = 0{,}72 \,.\, 10^{-14}$. L'erreur totale sur (54) ne dépasse donc pas $(0{,}9 + 0{,}72)10^{-14}$ mais il faut y ajouter $0{,}15 \,.\, 10^{-14}$ pour l'arrondi d'addition et l'arrondi sur $\pi/6$. L'erreur absolue totale devient ainsi $1{,}77 \,.\, 10^{-14}$ et l'erreur relative s'obtient en divisant par arc tg h, d'où $6{,}8 \,.\, 10^{-14}$.

(*c*) Si $1 < x < 1/h$, le calcul de $1/x$ amène une erreur relative 10^{-14}, donc une erreur absolue $10^{-14}/x$ sur $1/x$, donc une erreur héréditaire valant

$$\frac{x}{1+x^2} 10^{-14} \tag{61}$$

sur arc tg $1/x$. Celle-ci est maximum pour $x = 1$ et vaut $0{,}5 \,.\, 10^{-14}$. D'autre part l'erreur absolue d'arrondi dans le calcul de arc tg $1/x$ ne dépasse pas $0{,}72.10^{-14}$ comme au paragraphe précédent. L'erreur absolue totale est donc $(0{,}5 + 0{,}72)10^{-14} = 1{,}22 \,.\, 10^{-14}$. A cela s'ajoute $0{,}15 \,.\, 10^{-14}$ pour l'arrondi sur $\pi/2$ et l'arrondi de la soustraction (53), donnant $1{,}37 \,.\, 10^{-14}$. La plus petite valeur de arc tg$x$ dans l'intervalle considéré est $\pi/2$ — arc tg $h \triangleq 1{,}3$ et l'erreur relative ne dépasse pas $(1{,}37/1{,}3)10^{-14} = 1{,}05 \,.\, 10^{-14}$.

(*d*) Pour $x < 1/h$, l'erreur (61) est maximum pour $x = 1/h$ et vaut $0{,}25 \,.\, 10^{-14}$. Celle-ci s'ajoute à l'erreur absolue totale $1{,}77 \,.\, 10^{-14}$ du paragr. (*b*) pour donner $2{,}02 \,.\, 10^{-14}$ sur arc tg $1/x$. Ajoutant $0{,}15 \,.\, 10^{-14}$ comme au paragr. (*c*) l'erreur absolue sur arc tg x est $2{,}17 \,.\, 10^{-14}$. La plus petite valeur de arc tg x étant $\pi/4 \triangleq 0{,}79$, l'erreur relative ne dépasse pas $(2{,}17/0{,}79)10^{-14} = 2{,}76 \,.\, 10^{-14}$.

En conclusion on a le tableau suivant pour l'erreur relative maximum en unités 10^{-14} :

$x < h$	[illegible]
$h < x < 1$	[illegible]
$1 < x < 1/h$	1,05
$1/h < x$	2,76

Tout comme pour sin x, et pour les mêmes raisons, l'approximation arc tg $x = x$ est à utiliser pour $x < 10^{-8}$. On remplacera de même arc tg x par $\pi/2$ pour $x > 10^8$.

Parmi les diverses formules de contrôle possibles, celle qui conduit aux calculs les plus courts fait intervenir $z = \sin y$ où $y = \text{arc tg } x$, l'expression $(1 - z^2)x^2 - z^2$ devant en principe s'annuler. Si ϵ est l'erreur relative dans le calcul de arc tg et η celle dans le calcul d'un sinus, z est en fait remplacé par

$$(1 + \eta) \sin (1 + \epsilon)y = (1 + \eta + \epsilon y/x)z \tag{62}$$

et on trouve pour $(1 - z^2)x^2 - z^2$ la valeur $-2(\epsilon y + \eta x)x$. Pour x grand on a donc un mauvais contrôle puisque le terme en ϵ à contrôler est masqué par un terme ηx beaucoup plus élevé. On peut cependant limiter le contrôle au cas $x \leqslant 1$, puisque le cas opposé s'y ramène par une division (contrôlée par elle-même) et une prise de complément facile à contrôler séparément. Ceci admis, et compte tenu de ce que $y \leqslant x$ pour $x \leqslant 1$, la tolérance est inférieure à $2(\epsilon + \eta)x^2$, le coefficient x^2 étant essentiel pour les petits arguments. Compte tenu de la valeur $\epsilon = 6{,}8 \,.\, 10^{-14}$ (sans passage à l'inverse), de la valeur d'arrondi maximum $\eta = 6{,}58 \,.\, 10^{-14}$ dans le calcul du sinus, et d'un supplément de 3.10^{-14} pour les arrondis dans le premier membre (obtenu à partir d'une analyse détaillée que nous ne reproduisons pas) la formule à adopter pour la tolérance est

$$|(1 - z^2)x^2 - z^2| \leqslant 23{,}2 \,.\, 10^{-14}x^2. \tag{63}$$

10. Calcul des exponentielles.

10.1. *Calcul de* 10^x.

Pour x négatif, on passe au calcul de 10^{-x} et on prend l'inverse à la fin. Admettons donc x positif et inférieur à 49, sans quoi le résultat ne pourrait être noté dans la machine. Soient $x = ab, cde\ldots$ les chiffres successifs de x ramené à virgule fixe ; on a

$$10^x = 10^{10a+b}10^{0,cd\cdots} = 10^p 10^{0,cde\cdots}$$

et les deux premiers chiffres ab constituent directement l'exposant p du résultat. On continue par

$$10^{0,cde\cdots} = 10^{0\,c}10^{00,de\cdots} \tag{64}$$

et $10^{0,c}$ est lu dans un table (cfr table V). On est ainsi ramené au calcul de $y = 10^x$ dans l'intervalle $0 \leqslant x < 0{,}1$. Pour les raisons mentionnées au § 2, on calcule $z = 10^x - 1$ et on obtient ensuite y par $y = z + 1$.

10.2. *Calcul de* $10^x - 1$ *dans* $0 \leqslant x < 0{,}1$.

Désignons par $z(x)$ cette fonction et considérons les valeurs z_1 et z_2 qu'elle prend pour deux arguments x_1 et x_2. La formule d'addition des exposants donne

$$z(x_1 + x_2) = (z_1 + 1)z_2 + z_1. \tag{65}$$

On calcule $10^{0,0de\cdots} - 1$ à partir de $z_2 = 10^{0,0d} - 1$ lu dans une table (cfr table V) et de $z_1 = 10^{0,00e\cdots} - 1$ donné par le développement en série de $10^x - 1$ limité maintenant à l'intervalle $0 \leqslant x < 0{,}01$. On combine finalement les deux résultats par (65).

Posant $h = 0{,}01$, le développement de Maclaurin de $10^x = e^{\mu x} = e^{x_1}$ où $\mu = \log_e 10 = 2{,}303\ldots$, arrêté au terme en $x^6/6\,!$ donne une erreur absolue inférieure à

$$(\mu h)^7 e^h/7\,! = 6{,}8.10^{-16}. \tag{66}$$

Soustrayant un polynôme de Tchebycheff symétrique autour de $h/2$ de façon à supprimer le terme en $x_1^6 = (\mu x)^6$, donc le polynôme

$$\mu^6 \mathrm{T}_6(2x/h - 1)2^{-5}(h/2)^6/6\,! \tag{67}$$

on ajoute à (66) l'erreur

$$\mu^6 h^6 2^{-11}/6\,! = 10^{-16} \tag{68}$$

et on obtient un polynôme

$$p(x) = a_0 + a_1 + a_2x^2 + \ldots + a_5x^5 \tag{69}$$

représentant 10^x avec une erreur absolue inférieure à $7{,}8 \,.\, 10^{-16}$. Par le procédé du § 3.5., l'intégrarion en x donne

$$\int_0^x 10^x dx = (10^x - 1)/\mu \tag{70}$$

et le polynôme

$$10^x - 1 = b_1x + b_2x^2 + \ldots + b_6x^6 \tag{71}$$

avec

$$b_i = \mu a_{i-1}/i \tag{72}$$

est une approximation relative avec la même erreur. On trouvera ces coefficients, rebaptisés a_i, à la table V.

Pour $0 < x < 0{,}01$, l'erreur relative sur $10^x - 1$ est $\epsilon_1 = 2{,}5 \cdot 10^{-14}$ d'après le § 5.5.

Pour $0{,}01 < x < 0{,}1$, on applique (65) avec l'erreur ϵ_1 ci-dessus sur z_1. L'erreur relative sur z_2 provient de la table et peut atteindre $0{,}5 \cdot 10^{-15}$ divisé par la plus petite valeur de la table (0,12), donc $\epsilon_2 = 0{,}42 \cdot 10^{-14}$. Appliquant la méthode du § 5.2. à (65) avec la notation $\epsilon = 10^{-14}$ pour l'arrondi des opérations elles-mêmes, on trouve sur z l'erreur relative

$$\epsilon + \frac{(1 + z_1)z_2(\epsilon_2 + 2\epsilon) + \epsilon_1(1 + z_2)z_1}{z_1 + (z_1 + 1)z_2}.$$

La valeur de z_1, ne dépassant pas 0,023, est négligeable devant l'unité ; d'autre part $1 + z_2$ ne dépasse pas 1,23, de sorte que l'erreur relative ne dépasse pas

$$\epsilon + \frac{2{,}42 + 3{,}08 z_1/z_2}{1 + z_1/z_2}. \tag{73}$$

Comme z_1/z_2 varie de 0 à 1, le maximum se produit pour $z_1 = z_2$ et vaut $3{,}75 \cdot 10^{-14}$.

Pour $0 < x < 0{,}1$, l'erreur relative sur $10^x = y$ s'obtient en calculant $y = z + 1$ à partir de l'erreur relative $\epsilon_a = 3{,}75 \cdot 10^{-14}$ sur z et en ajoutant $\epsilon = 10^{-14}$ par opération. On trouve une erreur relative $\epsilon_a + \epsilon - \epsilon_a/y$ maximum pour $x = 0{,}1$ où $y = 1.26$. Le résultat est $1{,}8 \cdot 10^{-14}$.

Pour $x > 0{,}1$, la multiplication (64) ajoute l'erreur relative maximum de la table de $10^{0,c}$, soit $0{,}42 \cdot 10^{-14}$, et encore 10^{-14} pour l'arrondi de multiplication. L'erreur totale atteint donc $3{,}22 \cdot 10^{-14}$.

Pour $x < 0$, le passage à l'inverse ajoute 10^{-14} partout.

10.3. *Programme pour le calcul de* $z = 10^x - 1$.

Pour $|x| > 0{,}1$ on calcule $y = 10^x$ selon le programme du § 10.1, suivi de celui du § 10.2. On achève par $z = y - 1$, ce qui transforme une erreur relative ϵ sur y en une erreur relative $\epsilon(z + 1)/z$ sur z, qui est maximum pour $z = 10^{0,1} - 1 = 0{,}26$

et vaut $4{,}85\epsilon$. Comme, en unités 10^{-14}, ϵ valait 3,22 ou 4,222, selon le signe de x, on obtient ici 15,6 ou 20,5.

Pour $|x| < 0{,}1$ on distingue deux cas selon le signe de x. Pour x positif, on suit directement la méthode du § 10.2. Pour x négatif, on pose $x' = -x$ et on calcule $z' = 10^{x'} - 1$ selon le § 10.2. On repasse ensuite à $z = 10^x - 1$ par

$$z = \frac{-z'}{1 + z'} \tag{74}$$

L'erreur relative sur z' est au maximum $3{,}75 \,.\, 10^{-14}$ d'après le calcul suivant (73) ; elle se transmet héréditairement sur z certainement sans accroissement. D'autre part les trois opérations (addition, inversion, multiplication) intervenant dans (74) introduisent au maximum un arrondi $3 \,.\, 10^{-14}$, ce qui donne une erreur totale de $6{,}75 \,.\, 10^{-14}$.

10.4. *Contrôles.*

Le tableau ci-dessous résume les erreurs relatives (en unités 10^{-14}) obtenues dans les divers cas.

	10^x	$10^x - 1$
$0{,}1 < x$	3,22	15,6
$0 < x < 0{,}1$	1,8	3,75
$-0{,}1 < x < 0$	2,8	6,75
$x < -0{,}1$	4,22	20,5

Posant $y_1 = 10^{0{,}5x}$ et $z_1 = y_1 - 1$, le contrôle se fera en calculant $y - y_1^2$ ou $z - z_1(z_1 + 2)$ qui doivent théoriquement s'annuler. Dans le cas de y, on peut faire le contrôle avant le calcul éventuel de l'inverse puisque ce dernier possède son propre contrôle. On reste alors dans le cas de $x > 0$ où l'erreur relative ne dépasse pas $3{,}22 \,.\, 10^{-14}$. Cette erreur est doublée dans y_1^2 ce qui donne une erreur triple sur $y - y_1^2$. Par suite de l'arrondi dans y_1^2 (la soustraction entre nombres voisins n'en produit pas), il convient d'ajouter 10^{-14} et d'adopter finalement

$$|y - y_1^2| \leqslant 10{,}66 \,.\, 10^{-14} y. \tag{75}$$

Pour z, le calcul est ramené à celui de y dans le cas $|x| > 0{,}1$ et le contrôle sur y suffit puisque la seule opération restante est la soustraction de l'unité. Dans le cas $|x| \leqslant 0{,}1$, on doit également contrôler le cas de $x < 0$ puisqu'il ne s'agit pas d'une simple inversion. Un raisonnement analogue au précédent mène à une tolérance relative valant $(3 \times 6{,}75 + 1)10^{-14}$, donc

$$|z - z_1(z_1 + 2)| \leqslant 21{,}2 \,.\, 10^{-14} z. \tag{76}$$

11. Calcul des logarithmes.

11.1. *Préliminaires.*

Pour connaître la fonction $\log_{10}(1 + y)$ avec une erreur relative de l'ordre de 10^{-14} il faut un développement direct de cette fonction pour les petits arguments y tant positifs que négatifs, donc une série valant pour $|y| \leqslant h$. Pour $|y| > h$, on peut poser $1 + y = x$ et ramener le calcul à celui de $\log_{10} x'$, où x' est un nouvel argument compris entre $1 - h$ et $1 + h$, par une formule du type

$$\log x = \log z + \log x' \tag{77}$$

avec $x' = x/z$, où $\log x'$ se calcule de nouveau par développement en série, et où $\log z$ varie par pas discrets selon l'intervalle où se trouve x. Pour que le même développement en série vale dans tous les cas, il faut $1 - h < x' < 1 + h$, et (77) vaut pour une valeur donnée pour z dans un intervalle $x_a < x < x_b$ tel que

$$x_a/z = 1 - h \;;\;\; x_b/z = 1 + h,$$

donc

$$z = (x_a + x_b)/2 \tag{78}$$

et

$$x_b/x_a = a = \frac{1 + h}{1 - h}\,. \tag{79}$$

Partant de l'extrémité de $x_a = 1 + h$ de l'intervalle de base, on trouve un premier intervalle allant jusqu'à $a(1 + h)$, un second jusque $a^2(1 + h)$, etc. et, pour le i-ème intervalle, on a

$$x_b = a^i(1 + h) \;;\; x_a = a^{i-1}(1 + h) \tag{80}$$

et, par (78) et (79)

$$z = a^i(1 + 1/a)(1 + h)/2 = a^i. \qquad (81)$$

Il est commode de choisir n intervalles de telle façon que la dernière valeur de z soit 10, ce qui exige

$$a^n = 10. \qquad (82)$$

Il est en effet facile dans ce cas de reconnaître l'intervalle i dans lequel se trouve un nombre x compris entre $1 + h$ et 10 : cet intervalle est défini par

$$a^{i-1}(1 + h) < x < a^i(1 + h).$$

Divisant par $(1 + h)$, élevant à la n-^{ieme} puissance, et tenant compte de (82), il vient

$$10^{i-1} < u^n < 10^i\,; \quad u = \frac{x}{1 + h} \qquad (83)$$

et on obtient i en calculant u^n, changeant son exposant en mantisse, et ajoutant l'unité.

Il reste à déterminer les meilleures valeurs à adopter pour h et n (liés par (82) et (79)). Ce choix est régi par les considérations suivantes :

(*a*) Il faut que h soit assez petit pour que le développement en série dans l'intervalle de base n'exige pas trop de termes.

(*b*) Il faut que h soit assez grand pour que, pour x immédiatement supérieur à $1 + h$, argument pour lequel (77) intervient et implique une soustraction, il n'y ait pas de chiffres significatifs perdus. Ceci exige pratiquement que la valeur $\log_{10}(1 + h) = 0{,}43h$ à calculer ne soit pas franchement inférieure à 0,1. Un bon ordre de grandeur est donc $h = 0{,}23$. En fait un tel intervalle demande encore beaucoup de termes dans le développement, et nous avons légèrement sacrifié la précision en prenant $h = 0{,}15$, ce qui donne par (79) $a = 1{,}35$...

(*c*) La détermination de i par (83) exige le calcul d'une n-ième puissance, donc un certain nombre de multiplications. Pour réduire ce nombre, il y a avantage à prendre pour n une puissance exacte de 2, de façon à n'avoir que des élévations successives au carré. En prenant $n = 8$, (82) donne $\log_{10} a = 0{,}125$ et $a = 1{,}333 \ldots$, ce qui est inférieur à la valeur adoptée en (b) ; chacun des huit intervalles sera donc légèrement inférieur à l'intervalle de base, et le développement en série y sera valide a fortiori.

En conclusion, pour $1{,}15 < x < 10$, on détermine i par (83) avec $u = gx$, où

$$g = \frac{1}{1+h} = 0{,}869\ldots$$

(valeur exacte en table VI), et on a

$$\log_{10} x = im + \log_{10} x' \tag{84}$$

où

$$m = \log_{10} a = 0{,}125$$

et

$$x' = x/z = xw_i \tag{85}$$

les valeurs de $w_i = a^{-i}$ étant données à la table VI, à mémoriser dans la machine pour éviter des opérations supplémentaires.

11.2. *Programme de* $\log_{10}$ $(1 + y)$.

Si $|y| > 0{,}15$ on pose $x = 1 + y$ et on passe au programme du § 11.3. Si $|y| \leqslant 0{,}15$, on passe au développement en série du § 11.4.

11.3. *Programme de* $\log_{10} x$.

Après avoir contrôlé que x est positif, on pose $x = ,abcd\ldots 10^p$ et $x_1 = ,abcd\ldots 10^1$ où $1 < x_1 < 10$. On a alors $\log_{10} x = (p - 1) + \log_{10} x_1$. Si $x_1 < 1{,}15$ on pose $y_1 = x_1 - 1$ et on passe au développement en série du § 11.4. Sinon, on applique (83) et (84) et on est ramené au calcul de $\log x_1'$. En posant alors $y_1' = x_1' - 1$ on a de nouveau le programme du § 11.4. avec $|y_1'| < 0{,}15$.

11.4. *Développement en série.*

Le développement de $\log_{10}(1 + y) = \log_e(1 + y)$ avec $\mu = \log_{10} e = 0{,}434\ldots$ dans l'intervalle $|y| \leqslant h$ avec $h = 0{,}15$ s'obtient par le procédé du § 3.5. à partir de (10), où $\theta = y/h$, qu'il faut ensuite intégrer de 0 à y et multiplier par μ. L'erreur relative est (19), avec $r = 0{,}0754\ldots$ (donné par (18)), soit $e = 0{,}47 \cdot 10^{-15}$ pour $n = 13$, donc un polynôme $a_1 y + \ldots + a_{14} y^{14}$ après intégration, dont les coefficients sont donnés à la table VI.

11.5. *Erreur relative sur* $\log_{10}(1 + y)$.

Pour $|y| < 0{,}15$, on sait d'après le § 5.5. que l'erreur relative sur $\log_{10}(1 + y)$ ne dépasse pas $2{,}5 \cdot 10^{-14}$.

Pour $|y| > 0{,}15$, le passage à $x = 1 + y$ introduit un arrondi relatif de 10^{-14}, qui se transmet héréditairement à x' par (85), et l'arrondi de la table de w_i, et l'arrondi de multiplication, portent l'erreur relative sur x' à $2{,}5 \cdot 10^{-14}$. Elle donne héréditairement une erreur absolue $0{,}43 \times 2{,}5 \cdot 10^{-14}$ sur $\log_{10} x'$, donc $1{,}08 \cdot 10^{-14}$. D'autre part x' étant ramené à l'intervalle de base on calcule $y' = x' - 1$, ce qui donne un arrondi 10^{-14} sur y', se transmettant héréditairement sur $\log (1 + y')$ pour donner une erreur absolue inférieure à $0{,}43 \cdot 10^{-14}$. En troisième lieu l'erreur de calcul du logarithme produit un arrondi relatif de $2{,}5 \cdot 10^{-14}$, donc une erreur absolue $(2{,}5 \cdot 10^{-14})$ $(0{,}43\ h)$; pour $h = 0{,}15$ ceci vaut $0{,}17 \cdot 10^{-14}$. L'erreur absolue totale sur $\log_{10} x'$ est donc $(1{,}08 + 0{,}43 + 0{,}17) 10^{-14} = 1{,}68 . 10^{-14}$. Mais, compte tenu de toutes réductions, on a

$$\log_{10} x = p - 1 + im + \log_{10} x'. \tag{86}$$

Pour $1/10 < x < 10$, on a $|p - 1 + im| < 1$, et cette quantité est connue exactement par suite de la valeur ronde de m. Dans ce cas l'arrondi absolu d'addition dans (86) est 10^{-15} et l'erreur absolue sur $\log_{10} x$ est $1.78 . 10^{-14}$. Comme la plus petite valeur de $\log_{10} x$ est $\log_{10} 1{,}15 = 0{,}066$, l'erreur relative ne dépasse pas $27 \cdot 10^{-14}$. Pour $|p - 1 + im| > 1$, l'arrondi d'addition dans (86) peut augmenter d'un ou deux ordres de grandeur, le cas le plus défavorable en erreur relative se produisant pour $p = 11$, de façon que la constante dans (86) dépasse tout juste 10 ; l'arrondi est alors 10^{-13} et l'erreur absolue atteint $11{,}7 \cdot 10^{-13}$. Comme $\log_{10} x$ vaut au moins 10 dans ce cas, l'erreur relative est cependant inférieure à celle calculée ci-dessus.

11.6. *Erreur absolue sur* $\log_{10}x$ *et contrôle.*

Pour $\log_{10}x$, c'est l'erreur absolue qui compte. En outre, du fait que x est donné, l'erreur sur x' calculée au § 11.5 baisse de $2{,}5 \cdot 10^{-14}$ à $1{,}5 \cdot 10^{-14}$ et l'erreur sur $\log_{10}x'$ baisse de $1{,}68 \cdot 10^{-14}$ à $1{,}31 \cdot 10^{-14}$. Gardant 10^{-13} comme plus grand arrondi, l'erreur absolue sur $\log_{10}x$ ne dépasse pas $11{,}3 \cdot 10^{-13}$.

Dans la formule de contrôle

$$|\log_{10}x - \log_{10}x/2 - \log_{10}2| \leqslant \eta \tag{87}$$

$x/2$ est affecté d'une erreur relative 10^{-14} qui devient une erreur absolue $0{,}43 \cdot 10^{-14}$ sur son logarithme. Ajoutant $2 \times 11{,}3 \cdot 10^{-13}$ pour les deux logarithmes, $0{,}5 \cdot 10^{-15}$ pour l'arrondi de log 2, on voit qu'il faut $\eta = 22{,}65 \cdot 10^{-13}$.

11.7. *Contrôle sur* $\log_{10}(1 + y)$.

Pour $|y| > 0{,}15$, (87) convient. Au contraire ce contrôle est illusoire aux petits arguments. On emploiera $\log(1 + y)^2 = 2\log(1 + y)$ en calculant par $(1 + y)^2 = 1 + y'$ avec $y' = (2 + y)y$. Dans $\log(1 + y)$, l'erreur relative est $2{,}5 \cdot 10^{-14}$ et passe à $3{,}5 \cdot 10^{-14}$ après arrondi de la multiplication par 2, ce qui donne une erreur absolue $(3{,}5 \cdot 10^{-14})(0{,}43y) = 1{,}6 \cdot 10^{-14}y$. Les deux opérations intervenant dans le calcul de y' donnent un arrondi 2.10^{-14} en erreur relative, qui induit une erreur relative plus faible sur $\log(1 + y')$. D'autre part le calcul lui-même de $\log(1 + y')$ produit une erreur relative de $27 \cdot 10^{-14}$ (car y' peut dépasser 0,15), et l'erreur relative totale sur $\log(1 + y')$ est $29 \cdot 10^{-14}$, ce qui donne une erreur absolue $(29 \cdot 10^{-14})(0{,}43y') = 12{,}5 \cdot 10^{-14}y' = 25 \cdot 10^{-14}y$. La tolérance est donc

$$|\log_{10}(1 + y') - 2\log_{10}(1 + y)| \leqslant 26{,}6 \cdot 10^{-14}|y|. \tag{88}$$

12. Constantes numériques intervenant dans les calculs.

Les tables ci-dessous donnent ces constantes pour chacune des fonctions élémentaires. La notation des nombres est toujours de la forme (1) expliquée au § 1. Lorsque moins de 15 chiffres sont cités pour la mantisse, c'est que les derniers chiffres sont non pertinents. Les calculs numériques ont été effectués en partie par MM. R. Broeckx et F. Servais.

TABLE I. — *Constantes pour le calcul de* 1/x.

a_0	— 49689 441	+ 01
b_0	+ 58658 3851	+ 01
a_1	+ 21645 7031	+ 01
a_2	+ 20560 69	+ 01
a_3	+ 20097 3038	+ 01

TABLE II. — *Constantes pour le calcul de* $1/\sqrt{x}$.

a	+ 33183 6	+ 01
b	— 50545 7	+ 01
c	+ 33030 1	+ 01
$\sqrt{10}$	+ 31622 77660 16838	+ 01
a_1	+ 15072 87	+ 01

TABLE III. — *Constantes pour le calcul de sin* x.

$1/2\pi$	+ 15915 49430 91895	+ 00
a_1	+ 62831 85307 17958	+ 01
a_3	— 41341 70224 00904	+ 02
a_5	+ 81605 24927 36527	+ 02
a_7	— 76705 85964 78412	+ 02
a_9	+ 42058 69380 479	+ 02
a_{11}	— 15094 47101 00	+ 02
a_{13}	+ 38149 97692	+ 01
a_{15}	— 69993 87	+ 00

TABLE IV. — *Constantes pour le calcul de arc tg* x.

$\pi/2$	+ 15707 96326 79490	+ 01
h	+ 26794 91924 31123	+ 00
a	+ 57735 02691 89626	+ 00
$\pi/6$	+ 52359 87755 98299	+ 00
a_1	+ 10000 00000 00000	+ 01
a_3	— 33333 33333 33133	+ 00
a_5	+ 19999 99999 53049	+ 00
a_7	— 14285 71386 29367	+ 00
a_9	+ 11111 08297 21135	+ 00
a_{11}	— 90909 04310 57094	— 01
a_{13}	+ 76738 65538 37176	— 01
a_{15}	— 64383 08381 36175	— 01
a_{17}	+ 42953 79023 74788	— 01

TABLE V. — *Constantes pour le calcul de* $10^x - 1$ *et* 10^x.

$10^{0,1}$	+ 12589 25411 79417	+ 01
$10^{0,2}$	+ 15848 93192 46111	+ 01
$10^{0,3}$	+ 19952 62314 96888	+ 01
$10^{0,4}$	+ 25118 86431 50958	+ 01
$10^{0,5}$	+ 31622 77660 16838	+ 01
$10^{0,6}$	+ 39810 71705 53497	+ 01
$10^{0,7}$	+ 50118 72336 27273	+ 01
$10^{0,8}$	+ 63095 73444 80193	+ 01
$10^{0,9}$	+ 79432 82347 24282	+ 01
$10^{0,01} - 1$	+ 23292 99228 07541	— 01
$10^{0,02} - 1$	+ 47128 54805 08996	— 01
$10^{0,03} - 1$	+ 71519 30523 76094	— 01
$10^{0,04} - 1$	+ 96478 19614 31850	— 01
$10^{0,05} - 1$	+ 12201 84543 01963	+ 00
$10^{0,06} - 1$	+ 14815 36214 96883	+ 00
$10^{0,07} - 1$	+ 17489 75549 39529	+ 00
$10^{0,08} - 1$	+ 20226 44346 17413	+ 00
$10^{0,09} - 1$	+ 23026 87708 12381	+ 00
a_1	+ 23025 85092 99405	+ 01
a_2	+ 26509 49055 24004	+ 01
a_3	+ 20346 78591 64184	+ 01
a_4	+ 11712 55357 43595	+ 01
a_5	+ 53935 07360 70584	+ 00
a_6	+ 20697 89775 65653	+ 00

TABLE VI. — *Constantes pour le calcul des logarithmes.*

x	+ 86926 32173 91304	+ 00
m	+ 12500 00000 00000	+ 00
m_1	+ 13335 21432 16332	+ 01
m_2	+ 17782 79410 03892	+ 01
m_3	+ 23713 73705 66166	+ 01
m_4	+ 31622 77660 16838	+ 01
m_5	+ 42169 65034 28582	+ 01
m_6	+ 56234 13251 90349	+ 01
m_7	+ 74989 42093 32456	+ 01
m_8	+ 10000 00000 00000	+ 02
a_1	+ 43429 44819 03252	+ 00
a_2	— 21714 72409 51626	+ 00
a_3	+ 14476 48273 00840	+ 00
a_4	— 10857 36204 75287	+ 00
a_5	+ 86858 89644 76257	— 01
a_6	— 72382 41381 84136	— 01
a_7	+ 62042 64368 39413	— 01
a_8	— 54286 78368 15067	— 01
a_9	+ 48256 88263 78047	— 01
a_{10}	— 43431 74454 77172	— 01
a_{11}	+ 39378 66780 46232	— 01
a_{12}	— 36089 85953 34578	— 01
a_{13}	+ 36119 93229 58473	— 01
a_{14}	— 33077 46344 17417	— 01