Časopis pro pěstování matematiky a fysiky

J. M. Horák; Václav A. Hruška

Několik elementárních poznámek o numerickém počítání

Časopis pro pěstování matematiky a fysiky, Vol. 54 (1925), No. 1, 59--62

Persistent URL: http://dml.cz/dmlcz/123133

Terms of use:

© Union of Czech Mathematicians and Physicists, 1925

Institute of Mathematics of the Academy of Sciences of the Czech Republic provides access to digitized documents strictly for personal use. Each copy of any part of this document must contain these *Terms of use*.

This paper has been digitized, optimized for electronic delivery and stamped with digital signature within the project *DML-CZ: The Czech Digital Mathematics Library* http://project.dml.cz

Několik elementárních poznámek o numerickém počítání.

Napsali Dr. J. M. Horák a Dr. V. Hruška.

1. Definice. Řádem čísla nazývejme nejvyšší z řádů číslic, jimiž je vyjádřeno v desítkové soustavě číselné. Číslo a nazveme přibližnou hodnotou čili approximací čísla A, je-li rozdíl $v(a) = A - a$ co do prosté hodnoty menší než a i než A. Rozdílem $v(a)$ pak definujeme nepřesnost přibližné hodnoty a.*) Počtem přesných míst čili přesností čísla a nazýváme rozdíl $[a] = m - n$, značí-li m řád čísla a a n řád jeho nepřesnosti $v(a)$.**) Proto výroky: „Číslo má přesnost μ", „číslo má μ přesných míst" značí, že nepřesnost nedosahuje jedničky μ-tého místa od levého kraje čísla. Předpokládáme v dalším, že přibližné hodnoty mají jen tolik míst, aby nepřesnost byla menší, než jednotka posledního místa.

2. Přesnost součinu dvou čísel. Buďte m_1, m_2 řády přibližných hodnot a, b čísel A, B; n_1, n_2 buďte resp. řády jejich nepřesností, takže

$$(1) \qquad |v(a)| \leqq \varepsilon_1 \, 10^{n_1+1}, \qquad |v(b)| \leqq \varepsilon_2 \, 10^{n_2+1},$$

když ε_1, ε_2 značí vhodná čísla menší než 1. Označme ještě ε větší z obou čísel ε_1, ε_2 a

$$(2) \qquad a' = \frac{a}{10^{m_1}}, \qquad b' = \frac{b}{10^{m_2}}.$$

Čísla a', b' mají patrně týž číselný obraz jako a resp. b a o jejich velikosti platí

$$(3) \qquad 1 \leqq a' \leqq 10 - 10^{1-[a]}, \qquad 1 \leqq b' \leqq 10 - 10^{1-[b]}.$$

Patrně se můžeme omeziti na případ, že a, b a tedy též a', b' jsou kladná.

Nepřesnost součinu ab jest dána vzorcem

$$(4) \qquad v(ab) = AB - ab = a\,v(b) + (b + v(b))\,v(a).$$

Předpokládejme, že činitel a je nejméně tak přesný jako b, t. j. $[a] \geqq [b]$ a použijme vzorců (1) a (2). Dostaneme

$$(5) \qquad |v(ab)| < 10^{m_1+m_2+1-[b]}\,(a'.\varepsilon_2 + (b' + 10^{1-[b]})\,\varepsilon_1 . 10^{[b]-[a]}),$$

*) Franc. erreur, něm. Fehler.

**) Tato definice se liší poněkud od definice v Encyklop. d. math. Wiss. I., str. 981.

čili vzhledem k (3)

(6) $|v(ab)| < 10^{m_1+m_2+1-[b]}\,(a'.\varepsilon_2+\varepsilon_1)$ při $[a]>[b]$ a

(7) $|v(ab)| < 10^{m_1+m_2+1-[b]}\,(a'.\varepsilon_2+(b'+10^{1-[b]})\,\varepsilon_1)$ při $[a]=[b]$.

Vzhledem k (3) jest $\varepsilon_2 a'+\varepsilon_1 < 11\varepsilon$, takže při $[a]>[b]$ jest $|v(ab)|<11\varepsilon.10^{m_1+m_2+1-[b]}$. Jelikož součin ab je nejméně řádu m_1+m_2, nepřesnost nejvýše $1{\cdot}1\varepsilon$ jednotek řádu $m_1+m_2+2-[b]$, má součin aspoň $[b]-2$ přesných míst a nepřesnost jeho nedosáhne $1{\cdot}1\varepsilon$ jednotek místa $([b]-1)$-tého. Při $[a]=[b]$ jest možno psáti výhodně místo (7)

(8) $$|v(ab)| < 10^{m_1+m_2+1-[b]}.\varepsilon\,(a'+b'+10^{1-[b]}).$$

Výraz $a'+b'+10^{1-[b]}$ však jest menší vždy než 20, takže v případě, že ab je řádu m_1+m_2+1, má součin aspoň $[b]-1$ přesných míst, nepřesnost nedosahuje 2ε jednotek následujícího místa. Je-li však ab řádu m_1+m_2, bude $a'.b'<10$. Vyšetřme pro tento případ největší hodnotu, jíž nabude $a'+b'+10^{1-[b]}$. Pokud je $a'\geqq 1$, $b'\geqq 1$ (což jistě podle (3) je splněno), platí $a'b'=b'++b'\,(a'-1)\geqq b'+a'-1$, při čemž rovnost nastává jen pro $a'=1$ nebo $b'=1$. Má-li býti tedy $a'b'<10$, musí $a'+b'<11$; i jest pak také $a'+b'+10^{1-[b]}\leqq 11$, protože součet $a'+b'$ nemá cifry nižšího řádu než $10^{1-[b]}$ a nemůže se tedy od 11 lišiti o méně než $10^{1-[b]}$. Vidíme tedy, že je-li součin ab řádu m_1+m_2, opět nepřesnost součinu nedosáhne $1{\cdot}1\varepsilon$ jednotek místa $([b]-1)$.

Pravidlo 1. Jsou-li nepřesnosti obou činitelů menší než ε (<1) jednotek jejich posledních přesných míst, nedosahuje nepřesnost součinu nikdy $1{\cdot}1\varepsilon$ jednotek místa $(p-1)$-tého, je-li p počet přesných míst méně přesného činitele. V případě, že řád součinu je o jednu větší než součet řádů činitelů, jest tato horní hranice dokonce pouze $0{\cdot}2\varepsilon$ jedn. místa $(p-1)$-tého.

Je-li tedy $\varepsilon<\frac{10}{11}$ (specielně to je splněno při obvyklém braní oprav), bude součin míti $p-1$ přesných míst, nepřesnost nedosáhne jedné jednotky místa $(p-1)$-tého.

Pravidlo toto dává asi 9krát menší mez nepřesnosti než dosavadní pravidla podobného druhu.*) Osvědčuje se zvláště v případech, kdy nepřesnosti obou činitelů mají stejné znamení. Na př. $136{\cdot}7231.492{\cdot}64=67355{\cdot}267984$. Zkrátíme-li na $137.493=67541$, seznáváme, že skutečně nepřesnost $185{\cdot}732016$ jest menší, než $0{\cdot}36.1{\cdot}1=0{\cdot}396$ jednotek místa 2, t. j. než 396.

*) Encykl. cit., Vieille, Th. génér. des approx. numériques, Paris 1854. Fassbinder, Th. et prat. des approx. numériques, Paris 1906.

Přesnější odhad poskytují vzorce (6) a (7). Dosadíme na př. v (7) za a' a $b' + 10^{1-[b]}$ nejblíže vyšší celá čísla, t. j. $\alpha + 1$ a $\beta + 1$, jestliže α, β jsou nejvyšší číslice v obou činitelích. Vzorec ten nabude pak tvaru:

$$(10) \qquad |v(ab)| < 10^{m_1+m_2+1-[b]} \, (\varepsilon_2 \, (\alpha + 1) + \varepsilon_1 \, (\beta + 1)).$$

Při $[a] \neq [b]$ můžeme místo nejvyšší číslice činitele méně přesného položiti v (10) nulu (jde to z (6)). Podobně můžeme modifikovati tuto formuli, chceme-li přesnější mez v případě, že nepřesnosti obou faktorů mají různá znamení.

3 Dělení. Nepřesnost podílu bude

$$v\left(\frac{a}{b}\right) = \frac{a + v(a)}{b + v(b)} - \frac{a}{b} = \frac{b \, . \, v(a) - a \, . \, v(b)}{b\,(b + v(b))}.$$

Za označení přijatého v předešlém odstavci tedy je

$$(11) \qquad \left| v\left(\frac{a}{b}\right) \right| < 10^{m_1 - m_2 + 1 - p} \, . \, \frac{\varepsilon_2 \, a' \;\; 10^{p-[b]} + \varepsilon_1 \, b' \, . \, 10^{p-[a]}}{b' \;\; (b' - 10^{1-[b]})},$$

kde p značí menší z obou čísel $[a]$, $[b]$. Je-li $b' > 1$, t. j. $b > 10^{m_2}$, je výraz na pravé straně (11) menší než

$$10^{m_1-m_2+1-p} \; \varepsilon \; (10^{1+p-[b]} - 10^{1-[a]-[b]+p} + 10^{p-[a]}) < \\ < 11\varepsilon \, . \, 10^{m_1-m_2+1-p},$$

kterouž hodnotu jsme dostali dosazením $a' = 10 - 10^{1-[a]}$ a $b' = 1$, $b' - 10^{1-[b]} = 1$ do (11). Tudíž v případě, že podíl a/b je řádu $m_1 - m_2$, má aspoň $p - 2$ přesných míst, nepřesnost nedosahuje $1{\cdot}1\varepsilon$ jedn. místa $(p - 1)$. V případě, že a/b je řádu $m_1 - m_2 - 1$, vypadla by mez nepřesnosti 11ε jedn. místa $(p - 1)$, možno ji však zostřiti až na 2ε jednotek. V tomto případě totiž jest $a' < b'$, takže z (11) vyplyne

$$\left| v\left(\frac{a}{b}\right) \right| < 10^{m_1-m_2+1-p} \, . \, \varepsilon \, . \, 2,$$

když tam dosadíme $a'/b' < 1$, $b' - 10^{1-[b]} \geqq 1$. Nutno ještě vyšetřiti případ $b = 10^{m_2}$. Zlomek na pravé straně (11) nabude největší hodnoty při $a' = 10 - 10^{1-[a]}$ a to hodnoty

$$\frac{10^{1+p-[b]} - 10^{1+p-[a]-[b]} + 10^{p-[a]}}{1 - 10^{1-[b]}} \leqq 12\tfrac{1}{9},$$

je-li $[b] > 1$. Máme tedy:

Pravidlo 2. Nepřesnost podílu nedosáhne 2ε jednotek místa $(p - 1)$-tého, má-li méně přesný činitel (dělenec, dělitel) aspoň p přesných míst. Dokonce, je-li podíl řádu $m_1 - m_2$, klesne tato mez na $1{\cdot}1\varepsilon$, pokud není dělitel prostou mocninou desíti.

Také toto pravidlo podává mez nejméně 5krát, průměrně však aspoň desetkráte menší, než dosavadní pravidla (l. cit.).

Formuli vhodnou pro praktické počítání a platnou, pokud je $b \neq 10^{m_2}$, dostaneme z (11):

$$(12) \qquad \left| \nu\left(\frac{a}{b}\right) \right| < \varepsilon \cdot 10^{m_1 - m_2 + 1 - p} \frac{\alpha + \beta + 1}{\beta(\beta - 1)},$$

v níž dosazujeme i pro $\beta = 1$ za jmenovatel jednotku. Nejsou-li oba činitelé stejně přesní, můžeme klásti v čitateli jednotku místo první cifry méně přesného činitele.

4. Obě pravidla udávají obecnou mez nezávislou na velikosti přibližných hodnot. Z odvození je patrno, že tato mez může býti nejvýše asi desetkráte větší než hranice nepřesnosti získaná známým způsobem úvahou nepřesností relativních. Je zajímavo: kdyby základem soustavy bylo číslo g místo čísla 10 a kdybychom nepřesnost a počet přesných míst definovali jako v soustavě desítkové, mez nepřesnosti součinu bude $(1 + 1/g)\ \varepsilon$ jednotek místa $([b] - 1)$-tého a mez nepřesnosti podílu bude 2ε, tedy obecně počet přesných míst právě takový, jako v soustavě desítkové. Tyto meze jsou patrně nejmenší možné ze všech, které jsou tvaru $m \cdot \varepsilon$, kde značí m vhodné číslo a ε totéž co dříve.

*

Quelques théorèmes élémentaires sur le calcul numérique.

(Extrait de l'article précédent.)

Si l'on définit par $[a] = m - n$ le nombre de chiffres exacts d'une approximation a, m et n étant les ordres respectifs de a et de son erreur, on peut prononcer les théorèmes suivants, en y désignant par p le nombre de places exactes du facteur moins précis et en supposant qu'on ait supprimé tous les chiffres suivant le $(m - n)$-ième.

Théorème I. Si les deux nombres approchés ont leurs erreurs respectives inférieures à ε (< 1) unités de la dernière place, l'erreur de leur produit ne peut pas atteindre 1, $1\ \varepsilon$ unités de la $(p-1)$-*ième* place à partir du premier chiffre significatif à gauche. Si l'ordre du produit est supérieur à la somme des ordres des deux facteurs, on peut affirmer que cette limite est égale à 0, $2\ \varepsilon$ seulement.

Théorème II. L'erreur du quotient de deux nombres approchés ne peut pas atteindre 2ε unités de la place $(p-1)$-*ième* à partir du premier chiffre significatif à gauche, les notations employées ci-devant étant conservées. Si l'ordre du quotient est égal à la différence des ordres du dividende et du diviseur, on peut diminuer cette limite jusqu'à 1, $1\ \varepsilon$.

Ces limites sont les moindres possibles de celles qui ont la forme $m\varepsilon$, m étant une constante.

Il est bien intéressant de remarquer que ces limites deviennent $(1 + 1/g)\varepsilon$ et 2ε resp., si la base du système est g au lieu de 10.