Michal Křížek

Můžeme věřit numerickým výpočtům?

Pokroky matematiky, fyziky a astronomie, Vol. 56 (2011), No. 4, 290–297

Persistent URL: `http://dml.cz/dmlcz/142020`

Můžeme věřit numerickým výpočtům?

Věnováno Emilu Vitáskovi k jeho 80. narozeninám

Michal Křížek, Praha

Nikdy neztotožňujme realitu
s jejím matematickým
či numerickým modelem.

1. Úvod

Toto pojednání je volným pokračováním článku [6], jehož cílem bylo upozornit na mnohá skrytá nebezpečí při numerickém zpracování rozličných úloh bez znalosti teorie. Článků na podobné téma existuje celá řada – viz např. [10], [11], [12] či literatura uvedená v [6] nebo [16].

V úvodu klasické knížky *Numerical Processes in Differential Equations* [1] autorů Babušky, Prágera a Vitáska se uvádí několik špatně podmíněných úloh ukazujících na nestabilitu numerických výpočtů vzhledem k zaokrouhlovacím chybám. Např. výpočet integrálu

$$I_n = \frac{1}{\mathrm{e}} \int_0^1 x^n \mathrm{e}^x \,\mathrm{d}x > 0$$

lze metodou per partes převést na rekurenci

$$I_n = 1 - nI_{n-1}, \quad n = 1, 2, \dots,$$

kde $I_0 = 1 - \mathrm{e}^{-1}$. Pokud ji budeme vyhodnocovat na počítači v jednoduché aritmetice,[1] pak už po několika krocích dostaneme paradoxně zápornou hodnotu integrálů,[2] srov. též [2]. Je to způsobeno tím, že se v každém kroku odečítají dvě skoro stejně velká čísla. Při odečítání sice obecně nedochází k zaokrouhlování, ale mantisa rozdílu obsahuje jen málo platných míst, což nutně vede ke ztrátě přesnosti. Podobně je tomu při sčítání a odečítání dvou čísel s dosti odlišnými exponenty.

Jiný příklad v [1] se zabývá numerickou nestabilitou postupného provádění následujících operací

$$\dots(((((1 : 2) \cdot 2) : 3) \cdot 3) : 4) \cdot 4 \dots$$

[1]Kdysi měla jednoduchá aritmetika 4 byty, v současnosti má spíše 6 bytů (1 byte = 8 bitů). Terminologie ale není jednotná.

[2]Hodnoty I_n lze však velice přesně numericky vypočítat pomocí rychle konvergentní řady $I_n = 1/(n+1) - 1/((n+1)(n+2)) + 1/((n+1)(n+2)(n+3)) - \dots$

Prof. RNDr. Michal Křížek, DrSc., Matematický ústav AV ČR, v.v.i., Žitná 25, 115 67 Praha 1, e-mail: `krizekmath.cas.cz`

Po dostatečně mnoha krocích dostáváme na různých počítačích obecně zcela rozdílné výsledky.[3]

Cílem tohoto článku je upozornit na několik poměrně nových varovných příkladů, které by měl mít čtenář na paměti, pokud se bude pouštět do numerických výpočtů. Dva z nich jsou vybrány z knihy *Handbook of Floating-Point Arithmetic* [9], která nedávno vyšla. Jako šokující se jeví Mullerův příklad (viz kap. 5 dále), v němž se provádí 25 násobení a 25 odečítání s přesností na 25 platných míst, a výsledek je naprosto nesprávný. „Absolutní hausnumera“ dává také Rumpův příklad (viz kap. 6) s mnohem menším počtem prováděných aritmetických operací a na ještě větší počet platných míst.

2. Může být klesající posloupnost v počítačové aritmetice rostoucí?

Položme $a_0 = 1$, $a_1 = \frac{1}{11}$ a nechť

$$a_{n+2} = \frac{34}{11}a_{n+1} - \frac{3}{11}a_n \quad \text{pro } n = 0, 1, 2, \ldots \tag{1}$$

Přesné řešení této rekurence

$$a_n = \frac{1}{11^n}$$

je zřejmě klesající posloupnost. V počítačové aritmetice však z (1) dostáváme:

$a'_{50} \approx 2 \cdot 10^{11}$ ve zkrácené aritmetice (4 byty – short real),
$\tilde{a}_{50} \approx 10^{10}$ v jednoduché aritmetice (6 bytů – real),
$\hat{a}_{50} \approx 5 \cdot 10^5$ ve dvojnásobné aritmetice (8 bytů – long real),
$\overline{a}_{50} \approx 10^3$ v rozšířené aritmetice (10 bytů – extended).

Odpovídající posloupnosti (a'_n), $(\tilde{a}_n)$, $(\hat{a}_n)$ a $(\overline{a}_n)$ jsou (v tomto pořadí) rostoucí již od $n =$8, 9, 12, 14 a všechny dosti rychle rostou. Při zaokrouhlování jen na dvě platná místa je uvažovaná posloupnost rostoucí dokonce už od $n = 2$. Samozřejmě nejde o nekonečnou posloupnost, protože konečná aritmetika počítačů nám diktuje řadu omezení.

Pokud (1) implementujeme tak, že $a_{n+2} = (34a_{n+1} - 3a_n)/11$, pak výsledná posloupnost je dokonce záporná již od $\hat{a}_{12}$. Opět se tedy ukazuje, že odečítání dvou téměř stejně velkých čísel v počítačové (neceločíselné) aritmetice může způsobit řadu nepředvídatelných potíží.

3. Babuškův příklad

Na intervalu $[0, 1]$ uvažujme lebesgueovsky integrovatelnou funkci f definovanou vztahem

$$f(x) = \begin{cases} 1, & \text{je-li } x \text{ iracionální,} \\ 0, & \text{je-li } x \text{ racionální.} \end{cases}$$

Pak zřejmě

$$\int_0^1 f(x)\mathrm{d}x = 1,$$

[3] Tyto testy zveřejněné v monografii [1] prováděl Karel Segeth.

zatímco jakákoliv kvadraturní formule

$$\int_0^1 f(x)\mathrm{d}x \approx \sum_{i=1}^{n} w_i f(x_i)$$

s váhami $w_i \in \mathbb{R}$ a uzly $x_i \in [0,1]$ dává na počítači nulovou hodnotu integrálu, protože se uzly aproximují racionálními čísly zobrazitelnými v konečné počítačové aritmetice.

Tento absurdní výsledek, kdy chyba je rovna jedné, lze vysvětlit tím, že není splněn důležitý předpoklad hladkosti funkce f obsažený ve standardních větách o konvergenci kvadraturních formulí.

4. Kahanův příklad (viz [9])

Položme $a_0 = 2$, $a_1 = -4$ a uvažujme rekurenci

$$a_n = 111 - \frac{1130}{a_{n-1}} + \frac{3000}{a_{n-1}a_{n-2}} \qquad \text{pro } n = 2, 3, \ldots \tag{2}$$

Obecné řešení rekurence (2) má tvar

$$a_n = \frac{\alpha 100^{n+1} + \beta 6^{n+1} + \gamma 5^{n+1}}{\alpha 100^n + \beta 6^n + \gamma 5^n},$$

kde α, β a γ závisí na počátečních hodnotách a_0 a a_1.

Jestliže $\alpha \neq 0$, pak limita posloupnosti (a_n) je 100. Pokud ale $\alpha = 0$ a $\beta \neq 0$, limita se rovná 6. Pro $a_0 = 2$ a $a_1 = -4$ vychází $\alpha = 0$, $\beta = -3$ a $\gamma = 4$. Odtud plyne, že

$$a_{20} \doteq 6.034, \quad a_{30} \doteq 6.006,$$

ale ve dvojnásobné aritmetice z (2) dostaneme

$$\hat{a}_{20} \doteq 98.351, \quad \hat{a}_{30} \doteq 99.999$$

v důsledku zaokrouhlování.

5. Mullerův příklad (viz [9])

Nechť $a_1 = \mathrm{e} - 1 = 1.718281828\ldots$ a uvažujme naprosto nevinně vyhlížející posloupnost

$$a_n = n(a_{n-1} - 1) \qquad \text{pro } n = 2, 3, \ldots \tag{3}$$

Indukcí snadno odvodíme její explicitní tvar

$$a_n = n!\Big(\frac{1}{n!} + \frac{1}{(n+1)!} + \frac{1}{(n+2)!} + \cdots\Big). \tag{4}$$

Pro $n = 1$ totiž máme $a_1 = \mathrm{e} - 1$, a předpokládáme-li platnost vztahu (4) pro $n - 1 \geq 0$, dostaneme

$$n(a_{n-1} - 1) = n\Big((n-1)!\Big(\frac{1}{(n-1)!} + \frac{1}{n!} + \frac{1}{(n+1)!} + \cdots\Big) - \frac{(n-1)!}{(n-1)!}\Big) = a_n.$$

Z vyjádření (4) okamžitě vidíme, že posloupnost (a_n) je klesající a

$$\lim_{n\to\infty} a_n = 1.$$

Avšak po 25 odčítáních a 25 násobeních v (3) obdržíme dosti podivná čísla

$\tilde{a}_{25} \approx -1.306946321 \cdot 10^{13}$ v jednoduché aritmetice (6 bytů),
$\hat{a}_{25} \approx 1.201807248 \cdot 10^{9}$ ve dvojnásobné aritmetice (8 bytů),
$\overline{a}_{25} \approx -7.3557319606 \cdot 10^{6}$ v rozšířené aritmetice (10 bytů).

Rozšiřme tedy délku mantisy. To ale stále ještě nemusí vést k uspokojivým výsledkům. V aritmetice s D dekadickými ciframi totiž vychází prvních dvanáct cifer takto:

D	$a_{25}(D)$
20	615990.413139
21	– 4457.98859386
22	– 4457.98859386
23	195.374419140
24	40.2623187072
25	– 6.27131142281
26	1.48429359885

Teprve pro $D > 25$ se výsledky začínají podobat $a_{25} = 1.03993872967\ldots$. Např. pro $D = 26$ dostáváme shodu v první platné cifře a pro $D = 30$ dostaneme shodu prvních tří cifer za desetinnou tečkou, tj. $a_{25}(D) = 1.039897\ldots$. Všimněte si ještě, že hodnoty pro $D = 21$ a $D = 22$ jsou stejné, protože 21. desetinná cifra Eulerova čísla e je nula. Podobnou tabulku lze nalézt též ve vynikajícím popularizačním článku [10] E. Pelantové a M. Znojila, kde se uvádí i zajímavá aplikace posloupnosti (3) v bankovnictví.

Podívejme se nyní podrobněji na hlavní důvod tohoto podivného chování. Zaokrouhlovací chybu v i-tém kroku označme ε_i. Pak

$$\begin{aligned}
\tilde{a}_1 &= \mathrm{e} - 1 + \varepsilon_1 = a_1 + \varepsilon_1,\\
\tilde{a}_2 &= 2(\tilde{a}_1 - 1) + \varepsilon_2 = 2(a_1 + \varepsilon_1 - 1) + \varepsilon_2 = a_2 + 2!\varepsilon_1 + \varepsilon_2,\\
\tilde{a}_3 &= 3(\tilde{a}_2 - 1) + \varepsilon_3 = 3(a_2 + 2\varepsilon_1 + \varepsilon_2 - 1) + \varepsilon_3 = a_3 + 3!\varepsilon_1 + 3\varepsilon_2 + \varepsilon_3,\\
&\ \vdots\\
\tilde{a}_{25} &= 25(\tilde{a}_{24} - 1) + \varepsilon_{25} = a_{25} + 25!\varepsilon_1 + \frac{25!}{2!}\varepsilon_2 + \frac{25!}{3!}\varepsilon_3 + \cdots + \varepsilon_{25}.
\end{aligned}$$

Celková zaokrouhlovací chyba je tedy

$$a_{25} - \tilde{a}_{25} = -25!\Big(\varepsilon_1 + \frac{1}{2!}\varepsilon_2 + \frac{1}{3!}\varepsilon_3 + \cdots + \frac{1}{25!}\varepsilon_{25}\Big) \tag{5}$$

a její velikost podstatně závisí na několika prvních zaokrouhlovacích chybách $\varepsilon_1, \varepsilon_2, \ldots$

Tento rafinovaný příklad ukazuje, jak je důležité během numerického výpočtu sledovat, abychom od sebe neodečítali dvě téměř stejně velká čísla (srov. (3) a vztah (7) níže). Z vyjádření (5) je patrno, proč je výše uvedená rekurence (3) extrémně citlivá na zaokrouhlovací chyby.

6. Rumpův příklad (viz [13])

Mnohem menší počet odčítání než v předchozím příkladě může také vést ke zcela zkresleným numerickým výsledkům. Uvažujme racionální funkci

$$u(x,y) = 333.75y^6 + x^2(11x^2y^2 - y^6 - 121y^4 - 2) + 5.5y^8 + \frac{x}{2y}.$$

Pro $x = 77617.0$ a $y = 33096.0$ na bývalém počítači IBM 370 vyšlo[4]

$$\begin{aligned}
\tilde{u}(x,y) &= 1.172603 \text{ v jednoduché aritmetice,}\\
\hat{u}(x,y) &= 1.1726039400531 \text{ ve dvojnásobné aritmetice,}\\
\overline{u}(x,y) &= 1.172603940053178 \text{ v rozšířené aritmetice.}
\end{aligned}$$

Nad těmito výsledky bychom však neměli jásat.[5] Jednoduchá aritmetika nám sice dává velice pěknou shodu s dvojnásobnou i rozšířenou aritmetikou, ale přesná hodnota je pro $x = 77617.0$ a $y = 33096.0$ záporná

$$u(x,y) = -\frac{54767}{66192} = -0.827396059946821\ldots \tag{6}$$

Víte, proč se dospělo k tak podivnému výsledku? Hlavní důvod je ten, že při provádění aritmetických operací s obrovskými čísly dochází v průběhu výpočtu ke ztrátě přesnosti některých důležitých cifer. Všimněte si např., že rozdíl čísel $x^2 = 6024398689$ a $5.5y^2 = 6024398688$ je jedna. Při výpočtu $-x^2y^6 + 5.5y^8$ se tedy opět odečítají dvě téměř stejně velká čísla.

Detailní analýza Rumpova paradoxního příkladu je podána v [3] (viz též [8]). Položíme-li $z = 333.75y^6 + x^2(11x^2y^2 - y^6 - 121y^4 - 2)$ a $w = 5.5y^8$, pak

$$\begin{aligned}
z &= -7917111340668961361101134701524942850,\\
w &= +7917111340668961361101134701524942848.
\end{aligned}$$

Jinými slovy, z a w mají prvních 35 cifer stejných a liší se jen ve dvou posledních. Odtud plyne, že

$$z + w = -2,$$

a tudíž platí (6),

$$u(x,y) = -2 + \frac{x}{2y} = -\frac{54767}{66192}.$$

[4] Podobné výsledky dává i program MAPLE (verze 12) na běžném PC v aritmetice s D dekadickými ciframi pro $D = 7, 8, 9, 10, 12, 18, 26, 27$. Pro jiný počet cifer však vycházejí jiné hodnoty. Teprve pro $D \geq 37$ začínají být numerické výsledky blízko správné hodnoty (6).

[5] Jiné programy mohou dávat zcela odlišné hodnoty. Např. MATLAB ve dvojnásobné aritmetice dává úplně nesmyslný výsledek, $\hat{u}(x,y) = -1.1806 \cdot 10^{21}$ a teprve od $D \geq 29$ začínají být numerické výsledky přijatelné. Některé programy zase neohlásí „přetečení aritmetiky“, angl. overflow.

7. Problém tří těles

Při simulaci vývoje Sluneční soustavy na miliardy let dopředu či dozadu se numeričtí experimentátoři většinou nestarají o velikost chyby, které se přitom dopouštějí. Podívejme se proto, jaká nebezpečí na ně číhají, a do jaké míry lze věřit tomu, co se ve skutečnosti napočítá.

Pro jednoduchost uvažujme pouze problém tří těles o hmotnostech m_i $(i = 1, 2, 3)$, která na sebe vzájemně gravitačně působí. Jejich vektorové trajektorie r_i jsou podle newtonovské mechaniky popsány známou soustavou autonomních nelineárních diferenciálních rovnic

$$r_i'' = G\Big(\frac{m_j(r_j - r_i)}{|r_j - r_i|^3} + \frac{m_k(r_k - r_i)}{|r_k - r_i|^3}\Big) \tag{7}$$

pro $i, j, k \in \{1, 2, 3\}$, $i \neq j \neq k \neq i$, s danými počátečními podmínkami na polohu $r_i(0)$ a rychlost $r_i'(0)$, kde G je gravitační konstanta. Všimněte si, že se ve jmenovateli odečítají dvě skoro stejně velká čísla, pokud se alespoň dvě tělesa k sobě přiblíží, tj. např. pro $r_i \approx r_j$. Při numerickém řešení tohoto problému vzniká nejen disktretizační chyba, ale i zaokrouhlovací chyby. Příklady z předchozích kapitol by měly být dostatečným varováním.

Zásadním nedostatkem při matematickém modelování ale je, že se často ztotožňuje navrhovaný model s realitou. Např. soustava rovnic (7) představuje matematický model s nekonečnou rychlostí šíření gravitační interakce. Tato rychlost je však ve skutečnosti jistě konečná, a proto newtonovská mechanika nikdy nepopisuje realitu naprosto přesně.[6] Např. newtonovský model problému pěti těles umožňuje, aby se všechna tělesa dostala v konečném čase do nekonečna (viz [14]).

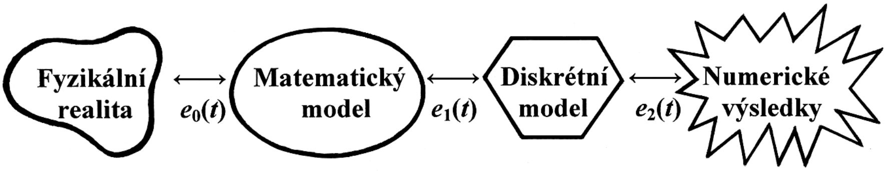
K chybám modelu dochází i při určování počátečních podmínek. Např. Merkur se při svém nejbližším přiblížení k Zemi posune o více než 3 své průměry, než jeho odražené světlo doputuje na Zemi. Když se jeho dráha počítá pomocí soustavy pro problém n těles, pak se jeho počáteční poloha určuje z toho, kde Merkur pozorujeme, a nikoliv, kde se ve skutečnosti právě nachází, což vyžaduje předpokládaná nekonečná rychlost šíření gravitační interakce. Vzniklá odchylka činí přes půl úhlové minuty!

8. Celková chyba aproximace

Na obrázku 1 je obecné schéma aproximace nějaké fyzikální reality na počítači, kde se vždy dopouštíme tří základních chyb e_0, e_1 a e_2, které se současně snažíme udělat co nejmenší. Pokud by však jedna z nich byla velká, pak i celková chyba $e = e_0 + e_1 + e_2$ bude velká. Více velkých chyb se obecně neruší. Navíc nemá velký smysl snažit se mít např. diskretizační chybu velice malou, když ostatní chyby budou velké.

Fyzikální realita se nahrazuje matematickým modelem, který je popsán např. soustavou diferenciálních rovnic (7). Vzniklá chyba modelu $e_0 = e_0(t)$ pak může být extrémně velká zejména pro dlouhé časové škály (viz [4] a [5]).

[6] Velké chyby modelu se tak dopouštíme, když aplikujeme newtonovskou mechaniku na simulaci vývoje galaxií v zakřiveném prostoročasu.



Obr. 1. Chyba modelu $e_0(t)$ je rozdíl mezi fyzikální realitou a jejím matematickým modelem. Diskrétní (nazývaný někdy též diskretizovaný) model se liší od matematického modelu o diskretizační chybu $e_1(t)$. Konečně v $e_2(t)$ jsou zahrnuty zaokrouhlovací chyby (ev. iterační chyba).

Již H. Poincaré věděl, že řešení soustavy (7) lze analyticky napsat jen v několika speciálních případech, ale obecně jej nelze explicitně vyjádřit nějakým vzorcem. Proto se hledá řešení přibližné. Nekonečněrozměrný matematický model se aproximuje konečněrozměrným diskrétním (někdy též nazývaným diskretizovaným) modelem, který lze implementovat na počítači. V případě problému tří těles se obvykle jedná o nějakou numerickou metodu pro řešení soustavy (7). Vzniklá diskretizační chyba e_1 je na dlouhých škálách také velká, protože soustava (7) není ljapunovsky stabilní, tj. malé změny v počátečních podmínkách způsobují velké změny v řešení.

K tomu, abychom zjistili, že diskretizační chyba je malá, se obvykle numerické řešení srovnává s řešením s poloviční délkou integračního kroku, o němž se předpokládá, že je „přesné“. Lepší ale je používat integraci vpřed a pak vzad, což nám díky jednoznačnosti “nekolizního” spojitého řešení soustavy (7) umožňuje stanovit, jak daleko jsme se vzdálili od původních počátečních podmínek.

O ztrátě přesnosti v důsledku zaokrouhlovacích chyb e_2 jsme si toho již hodně řekli v předchozích kapitolách.

9. Závěr

Archimedes počítal přibližně obsah jednotkového kruhu tak, že mu vepisoval a opisoval pravidelné mnohoúhelníky, jejichž obsahy uměl spočítat. Tím vlastně získal zaručené dvojstranné odhady chyby čísla π. Pomocí známé Lehmerovy-Schurovy metody zase můžeme stanovit libovolně malý kruh v komplexní rovině, ve kterém se nachází kořen daného polynomu. Podobné apriorní či aposteriorní odhady chyby se odvozují i pro řešení diferenciálních rovnic (viz např. [7], [15]). Pokud neprovedeme odhad chyby, nevíme ani, co jsme vlastně spočítali.

Operace sčítání a násobení nejsou asociativní v reálné aritmetice počítačů. Numerické výsledky navíc závisí na typu počítače, způsobu naprogramování apod. Správně sestavený program by měl ohlásit, zda v průběhu výpočtu nedošlo někde k odečtení dvou téměř stejně velkých čísel v reálné aritmetice.

Poděkování. Autor děkuje Ing. Ľ. Balkové, Ph.D., doc. RNDr. J. Chlebounovi, CSc., Ing. J. Mikyškovi, Ph.D., prof. RNDr. E. Pelantové, CSc., Mgr. A. Pravdové, Ph.D., RNDr. T. Vejchodskému, Ph.D., a prof. Z. Zhangovi, Ph.D., za inspirující diskuse. Článek byl podpořen výzkumným záměrem MSM 0021620839 a grantem IAA 100190803 GA AV ČR.

L i t e r a t u r a

[1] BABUŠKA, I., PRÁGER, M., VITÁSEK, E.: *Numerical processes in differential equations.* John Wiley & Sons, London, New York, Sydney 1966.

[2] BABUŠKOVÁ, R.: *Über numerische Stabilität einiger Rekursionsformeln.* Apl. Mat. *9* (1964), 186–193.

[3] CUYT, A., VERDONK, B., BECUWE, S., KUTERNA, P.: *A remarkable example of catastrophic cancellation unraveled.* Computing *66* (2001), 309–320.

[4] KŘÍŽEK, M.: *Does a gravitational aberration contribute to the accelerated expansion of the Universe?* Comm. Comput. Phys. *5* (2009), 1030–1044.

[5] KŘÍŽEK, M.: *Dark energy and anthropic principle.* New Astronomy (2011), 1–30.

[6] KŘÍŽEK, M., PRÁGER, M., VITÁSEK, E.: *Spolehlivost numerických výpočtů.* PMFA *42* (1997), 8–23.

[7] KŘÍŽEK, M., ROOS, H.-G., CHEN, W.: *Two-sided bounds of the discretization error for finite elements.* ESAIM Math. Model. Numer. Anal. *45* (2011), 915–924.

[8] LOH, E., WALSTER, G. W.: *Rump's example revisited.* Reliable Comput. *8* (2002), 245–248.

[9] MULLER, J.-M., ET AL.: *Handbook of floating-point arithmetic.* Birkhäuser 2009.

[10] PELANTOVÁ, E., ZNOJIL, M.: *Můžeme věřit své vlastní kalkulačce?* Rozhledy mat.-fyz. *85* (2010), 11–18.

[11] POKORNÝ, P.: *Úskalí číslicových počítačů při simulaci nelineárních dynamických systémů.* Sborník semináře: Determinismus a chaos, Herbertov (ed. L. Herrmann), ČVUT, Praha (2005), 47–51.

[12] PRÁGER, M., SÝKOROVÁ, I.: *Jak počítače počítají.* PMFA *49* (2004), 32–45.

[13] RUMP, S. M.: *Algorithms for verified inclusions – theory and practice.* In: Reliability in Compution (ed. R. E. Moore), Academic Press, New York, 1988, 109–126.

[14] SAARI, D. G., XIA, Z.: *Do nekonečna v konečném čase.* PMFA *42* (1997), 90–102.

[15] VEJCHODSKÝ, T.: *Aposteriorní odhady chyby v metodě konečných prvků.* PMFA *53* (2008), 104–112.

[16] VITÁSEK, E., KŘÍŽEK, M.: *(Ne)spolehlivost numerických výpočtů. Jaká nebezpečí skrývá numerické počítání?* Sborník semináře: Programy a algoritmy numerické matematiky 9, Kořenov, MÚ AV ČR Praha, 1998, 139–150.