Numerické metody v inženýrských úlohách (YNMI)

Část II: Metoda konečných diferencí

Milan Jirásek, 3. 11. 2016

Úvod

Vzhledem k omezenému rozsahu se v tomto předmětu budeme zabývat hlavně obyčejnými diferenciálními rovnicemi (ODR) 1. řádu, které obvykle popisují vývoj nějakého systému v čase, takže nezávisle proměnnou označíme t a hledanou funkci budeme značit u. Pro ODR 1. řádu je třeba předpsat jednu podmínku, která má obvykle charakter počáteční podmínky v čase 0 a popisuje výchozí stav zkoumaného systému. Řešením rovnice pak získáme představu o dalším vývoji tohoto systému v časech t mezi 0 a T.

Pokud čas dovolí, zmíníme se také o ODR 2. řádu (s aplikací např. na popis kmitání mechanického systému s jedním stupněm volnosti) a o soustavách obyčejných diferenciálních rovnic.

1 Obyčejné diferenciální rovnice 1. řádu

1.1 Cauchyho úloha

Obecně by ODR 1. řádu mohla být zadána v implicitním tvaru

$$F(\dot{u}(t), u(t), t) = 0 \tag{1}$$

Pokud by ale pro některou kombinaci hodnot u a t tato rovnice byla splněna pro více hodnot $\dot{u}$, případně nebyla splněna pro žádné $\dot{u}$, nastaly by problémy — další vývoj systému, který se dostal do tohoto stavu, by nebyl jednoznačný, nebo by ho vůbec nebylo možné určit. Proto je rozumné předpokládat, že rovnici popisující reálnou inženýrskou úlohu lze formulovat rovnou v explicitním tvaru

$$\dot{u}(t) = f(u(t), t) \tag{2}$$

kde f je jistá funkce, která v závislosti na okamžitém stavu systému a na čase udává rychlost, jakou se stav mění. Nalezení funkce $u(t)$, která splňuje diferenciální rovnici (2) a počáteční podmínku

$$u(0) = \bar{u}_0 \tag{3}$$

se označuje jako Cauchyho úloha. Přitom $f(u, t)$ je daná funkce a $\bar{u}_0$ je daná počáteční hodnota. Úlohu (2)–(3) řešíme na konečném časovém intervalu $[0, T]$.

Na první pohled by se mohlo zdát, že Cauchyho úloha (2)–(3) má vždy jednoznačné řešení, protože výchozí hodnota u v čase 0 je dána počáteční podmínkou (3) a pro každý čas $t \geq 0$ je rovnicí (2) jednoznačně určeno, jak rychle se hodnota u mění. Podívejme se však na následující příklady zdánlivě "neškodných" rovnic.

Příklad A:

Řešte rovnici

$$\dot{u}(t) = u^2(t) \tag{4}$$

s počáteční podmínkou

$$u(0) = 1 \tag{5}$$

Při řešení můžeme využít tzv. separaci proměnných, tj. přepsat rovnici jako

$$\frac{\mathrm{d}u}{u^2} = \mathrm{d}t \tag{6}$$

a poté na obou stranách integrovat od počátečního do současného stavu. Tím dospějeme ke vztahu

$$-\frac{1}{u(t)} + 1 = t \tag{7}$$

a nalezneme řešení

$$u(t) = \frac{1}{1-t} \tag{8}$$

Problém je v tom, že toto řešení není definováno pro čas $t = 1$ a v okolí tohoto času se chová "divně" — exploduje do plus nekonečna a pak "přiběhne zpět" z mínus nekonečna. Vzhledem k nespojitosti v čase 1 si nemůžeme být jistí, že získané řešení (8) platí i pro $t > 1$, a tudíž nevíme, jak se po tomto čase bude náš systém chovat.

Příklad B:

Řešte rovnici

$$\dot{u}(t) = 2\sqrt{u(t)} \tag{9}$$

s počáteční podmínkou

$$u(0) = 0 \tag{10}$$

Separace proměnných vede k podmínce

$$2\sqrt{u(t)} = 2t \tag{11}$$

a tedy k řešení

$$u(t) = t^2 \tag{12}$$

V tomto případě řešení sice roste nade všechny meze, ale spojitě, a je definováno pro všechny časy $t \geq 0$. Zdánlivě je tedy vše v pořádku. Problém je ovšem v tom, že i funkce $u(t) = 0$ splňuje rovnici (9) a počáteční podmínku (10). Řešení dané Cauchyho úlohy tedy není jednoznačné a od samého počátku nevíme, jak se bude náš systém chovat.

Je tedy zřejmé, že pro obecný tvar pravé strany rovnice (2) nelze garantovat jednoznačnost řešení, ani jeho existenci pro všechny kladné časy. Proto se omezíme na případy, kdy funkce $f(u,t)$ je lipschitzovská vzhledem k proměnné u, což zhruba řečeno znamená, že při změně u se funkční hodnota může změnit nejvýš o jistý násobek změny u, jinými slovy, růst (či pokles) funkční hodnoty nemůže být libovolně rychlý. Přesná definice požaduje existenci takové reálné konstanty L, aby pro všechna (u_a, t) a (u_b, t) z definičního oboru funkce f platilo

$$|f(u_a, t) - f(u_b, t)| \leq L|u_a - u_b| \tag{13}$$

V úvahách o lokální chybě použijeme i silnější předpoklad, že funkce f je spojitě diferencovatelná.

1.2 Eulerova dopředná (explicitní) metoda

Přibližné numerické řešení budeme hledat pro časové okamžiky $t_0 = 0, t_1, t_2, \ldots t_{N-1}, t_N = T$. Symbolem u_n označíme aproximaci přesného řešení $u(t_n)$ v čase t_n, $n = 0, 1, 2, \ldots N$. Na základě počáteční podmínky (3) položíme $u_0 = \bar{u}_0$. Eulerova dopředná metoda aproximuje pravou stranu rovnice (2) v intervalu $[t_{n-1}, t_n]$ konstantou $f(u_{n-1}, t_{n-1})$. Odtud pak snadno vyplyne vztah pro výpočet přibližné hodnoty

$$u_n = u_{n-1} + h_n f(u_{n-1}, t_{n-1}), \qquad n = 1, 2, \ldots N \tag{14}$$

kde

$$h_n = t_n - t_{n-1}, \qquad n = 1, 2, \ldots N \tag{15}$$

je délka n-tého časového kroku. Tento postup je explicitní v tom smyslu, že vyžaduje pouze přímé vyhodnocení funkce f, nikoli řešení algebraických rovnic, jak je tomu u explicitní metody popsané v oddílu 1.3.

Příklad C:

Eulerovou dopřednou metodou sestrojte přibližné řešení rovnice

$$\dot{u}(t) = -u(t) \tag{16}$$

s počáteční podmínkou

$$u(0) = 1 \tag{17}$$

Tato úloha samozřejmě má jednoduché analytické řešení

$$u(t) = \mathrm{e}^{-t} \tag{18}$$

Znalosti přesného řešení můžeme využít k vyhodnocení chyby numerické metody.

Pokud aplikujeme Eulerovu dopřednou metodu s konstantním krokem h, získáme následující numerické řešení v časech h, $2h$, $3h$ atd.:

$$\begin{aligned}
u_1 &= u_0 + hf(u_0, 0) = u_0 - hu_0 = (1-h)u_0 = 1 - h && (19)\\
u_2 &= u_1 + hf(u_1, h) = u_1 - hu_1 = (1-h)u_1 = (1-h)^2 && (20)\\
u_3 &= \ldots = (1-h)^3 && (21)\\
&\ldots && (22)\\
u_N &= \ldots = (1-h)^N && (23)
\end{aligned}$$

Pro čas $T = Nh$ tedy získáváme přibližné řešení $u_N = (1-h)^N$ zatímco přesné řešení je $u(T) = \mathrm{e}^{-T} = \mathrm{e}^{-Nh}$. Zkoumejme nyní, jak pro pevně zvolený čas T závisí chyba na počtu dílků N, na který jsme rozdělili interval $[0, T]$. Tuto chybu můžeme vyjádřit jako

$$u(T) - u_N = \mathrm{e}^{-T} - \left(1 - \frac{T}{N}\right)^N \tag{24}$$

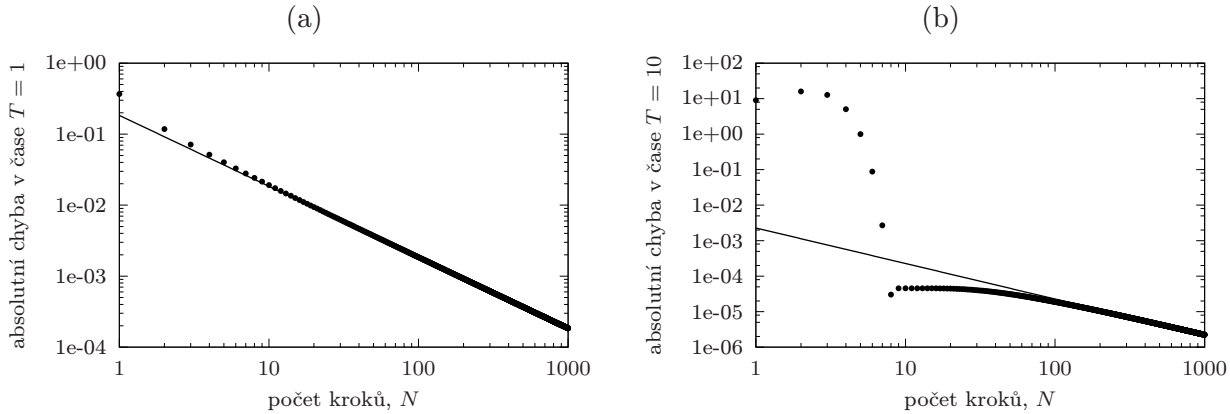
Připomeňte si známý vzorec

$$\lim_{k\to\infty} \left(1 + \frac{1}{k}\right)^k = \mathrm{e} \tag{25}$$

pomocí kterého můžeme ukázat, že

$$\lim_{N\to\infty}\left(1-\frac{T}{N}\right)^N = \mathrm{e}^{-T} \tag{26}$$

S rostoucím počtem kroků N (při pevně zvoleném T) tedy numerické řešení skutečně konverguje k přesné hodnotě. Otázka je, jak rychle. Chceme-li chybu snížit na desetinu, kolikrát menší krok máme použít?



Obrázek 1: Závislost absolutní chyby na počtu kroků: (a) chyba v čase $T = 1$, (b) chyba v čase $T = 10$

Základní představu můžeme získat vyhodnocením chyby pro zvolené T a rostoucí N. Na obr. 1 je závislost absolutní hodnoty chyby na počtu kroků N vynesena v logaritmickém měřítku pro dvě vybrané hodnoty času T. Jednotlivé body odpovídají skutečným chybám pro počty kroků dané celými čísly, zatímco přímky představují grafy funkce

$$e_T(N) = \frac{T^2}{2\mathrm{e}^T}\,\frac{1}{N} \tag{27}$$

která popisuje, jak se chyba vyvíjí asymptoticky pro $N \to \infty$.

Pro $T = 1$ se při zvyšování počtu kroků chyba monotónně zmenšuje, což odpovídá očekávání. Naproti tomu pro $T = 10$ je při nízkém počtu kroků vývoj chyby poněkud zvláštní, viz obr. 1b. Při výpočtu s jedním krokem má absolutní chyba hodnotu 10, která je o mnoho řádů větší než přesné řešení $\mathrm{e}^{-10} \approx 4{,}54 \times 10^{-5}$. Při výpočtu se dvěma nebo třemi kroky je dokonce ještě o něco větší a teprve poté začne klesat. Nicméně teprve pro 8 kroků se chyba dostane na úroveň srovnatelnou s přesným řešením a při dalším zvyšování počtu kroků zůstává zhruba konstantní. Poklesu chyby nepřímo úměrného počtu kroků N je dosaženo až pro N větší než 100.

Obrovské chyby při výpočtech s $T = 10$ a $N < 8$ souvisejí se ztrátou numerické stability, která bude podrobně analyzována později. Na tomto místě ještě ukažme, jak odvodit vztah (27) pro asymptotický vývoj chyby. Místo zkoumání limitního chování rozdílu $u(T) - u_N$ pro $N \to \infty$ je pohodlnější zavést proměnnou $x = 1/N$ a zkoumat limitní chování pro $x \to 0^+$. Pro dané T popíšeme chybu v závislosti na x funkcí

$$f(x) = \mathrm{e}^{-T} - (1-Tx)^{1/x} = \mathrm{e}^{-T} - \exp\left(\frac{1}{x}\ln(1-Tx)\right) \tag{28}$$

Pro $x \to 0^+$ se funkční hodnota blíží nule. Derivaci funkce f můžeme pro kladné x vyjádřit jako

$$f'(x) = \exp\left(\frac{1}{x}\ln(1-Tx)\right)\frac{Tx+(1-Tx)\ln(1-Tx)}{x^2(1-Tx)} \tag{29}$$

Exponenciální člen před zlomkem konverguje k e^{-T}. S využitím Taylorova rozvoje logaritmické funkce kolem bodu 1 vypočteme

$$\begin{aligned}\lim_{x\to 0^+}\frac{Tx+(1-Tx)\ln(1-Tx)}{x^2(1-Tx)} &= T^2\lim_{z\to 0^+}\frac{z+(1-z)\ln(1-z)}{z^2(1-z)} = \\ &= T^2\lim_{z\to 0^+}\frac{z-(1-z)(z+z^2/2+z^3/3+z^4/4+\ldots)}{z^2(1-z)} = \\ &= T^2\lim_{z\to 0^+}\frac{z^2/2+z^3/6+z^4/12+\ldots}{z^2(1-z)} = \frac{T^2}{2}\end{aligned} \tag{30}$$

Po dosazení do (29) dostaneme

$$\lim_{x\to 0^+} f'(x) = \mathrm{e}^{-T}\frac{T^2}{2} \tag{31}$$

takže pro $x \to 0^+$ máme

$$f(x) \approx \frac{T^2}{2\mathrm{e}^T}\,x \tag{32}$$

což znamená, že

$$u(T) - u_N \approx \frac{T^2}{2\mathrm{e}^T}\,\frac{1}{N} \tag{33}$$

Ze získaného odhadu vyplývá, že pro dostatečně vysoký počet kroků je chyba v pevně zvoleném čase T nepřímo úměrná počtu kroků N a tudíž přímo úměrná délce kroku $h = T/N$. Zkrácení kroku na polovinu pak vede také ke snížení chyby na polovinu. Hovoříme o lineární konvergenci metody.

Analýza chyby Eulerovy dopředné metody zde byla provedena na základě znalosti přesného řešení pro konkrétní Cauchyho úlohu (16)–(17). Podívejme se nyní, jak se odvozený výsledek dá zobecnit.

Lokální chyba

Zkoumejme numerické řešení obecné Cauchyho úlohy (2)–(3) Eulerovou dopřednou metodou. Kdybychom v $n-$tém časovém kroku vyšli z přesné hodnoty $u_{n-1} = u(t_{n-1})$, vznikla by na konci tohoto kroku tzv. lokální chyba ε_n, která je rozdílem přesného přírůstku

$$u(t_n) - u(t_{n-1}) = \int_{t_{n-1}}^{t_n} \dot{u}(t)\,\mathrm{d}t \tag{34}$$

a jeho aproximace (založené na přesné znalosti hodnoty na počátku kroku)

$$u_n - u(t_{n-1}) = h_n f(u(t_{n-1}), t_{n-1}) \tag{35}$$

Lokální chybu

$$\varepsilon_n = \int_{t_{n-1}}^{t_n} \dot{u}(t)\,\mathrm{d}t - h_n f(u(t_{n-1}), t_{n-1}) \tag{36}$$

lze odhadnout za předpokladu, že řešení $u(t)$ je dvakrát spojitě diferencovatelné[1] na intervalu $[0, T]$. Podle věty o střední hodnotě totiž existuje $\tau \in [t_{n-1}, t_n]$ takové, že

$$\int_{t_{n-1}}^{t_n} \dot{u}(t)\,\mathrm{d}t = (t_n - t_{n-1})\dot{u}(\tau) = h_n \dot{u}(\tau) \tag{38}$$

a po dosazení do (36) dostaneme

$$\varepsilon_n = h_n\left[\dot{u}(\tau) - f(u(t_{n-1}), t_{n-1})\right] = h_n\left[\dot{u}(\tau) - \dot{u}(t_{n-1})\right] \tag{39}$$

Přesnou hodnotu τ neznáme, ale víme, že se nachází v intervalu $[t_{n-1}, t_n]$, takže může být maximálně ve vzdálenosti h_n od t_{n-1}. Opětovným použitím věty o střední hodnotě dostaneme

$$\dot{u}(\tau) - \dot{u}(t_{n-1}) = \int_{t_{n-1}}^{\tau} \ddot{u}(t)\,\mathrm{d}t = (\tau - t_{n-1})\ddot{u}(\xi) \tag{40}$$

kde $\xi \in [t_{n-1}, \tau]$. Podle předpokladu je $\ddot{u}(t)$ spojitá funkce na uzavřeném intervalu, je tedy omezená a existuje C takové, že $|\ddot{u}(t)| \leq C$ pro všechna $t \in [0, T]$. Jelikož $|\tau - t_{n-1}| \leq h_n$, získáme z (40) odhad

$$|\dot{u}(\tau) - \dot{u}(t_{n-1})| \leq C h_n \tag{41}$$

a ten můžeme ve spojení s (39) využít k sestrojení odhadu

$$|\varepsilon_n| \leq C h_n^2 \tag{42}$$

Lokální chyba tudíž nemůže překročit jistý násobek druhé mocniny délky kroku. Zmenšíme-li krok desetkrát, dá se očekávat, že lokální chyba se zmenší stokrát. Jde však o chybu v rámci jednoho kroku, který je nyní mnohem kratší. Otázka je, jak se zmenšuje chyba aproximace $u(\bar{t})$ pro nějaké pevně zvolené $\bar{t}$, např. pro $\bar{t} = T$.

Globální chyba

V prvním kroku vycházíme z předepsané počáteční podmínky (3), takže po prvním kroku je rozdíl mezi přesným a přibližným řešením roven lokální chybě ε_1. Ve druhém kroku ale vznikne kromě lokální chyby ε_2 i chyba způsobená tím, že na jeho počátku už vycházíme z aproximace u_1 místo z přesné hodnoty $u(t_1)$. Nestačí ovšem pouze sečíst lokální chyby ε_1 a ε_2, protože chyba ε_1 vnesená do řešení na počátku kroku se může v průběhu kroku ještě zvětšit. To lze nejlépe vysvětlit, jestliže rozepíšeme celkovou chybu na konci druhého kroku následujícím způsobem:

$$\begin{aligned} u(t_2) - u_2 &= u(t_1) + \int_{t_1}^{t_2} \dot{u}(t)\,\mathrm{d}t - [u_1 + h_2 f(u_1, t_1)] = && (43)\\ &= u(t_1) - u_1 + && (44)\\ &\quad \int_{t_1}^{t_2} \dot{u}(t)\,\mathrm{d}t - h_2 f(u(t_1), t_1) + && (45)\\ &\quad h_2\left[f(u(t_1), t_1) - f(u_1, t_1)\right] && (46) \end{aligned}$$

[1]Druhá derivace řešení $u(t)$ je spojitá, jestliže jsou spojité první parciální derivace dané funkce $f(u, t)$. To plyne ze vztahu

$$\ddot{u}(t) = \frac{\partial f(u(t), t)}{\partial u}\dot{u}(t) + \frac{\partial f(u(t), t)}{\partial t} = \frac{\partial f(u(t), t)}{\partial u} f(u(t), t) + \frac{\partial f(u(t), t)}{\partial t} \tag{37}$$

který získáme diferenciací rovnice (2) podle času s uplatněním pravidla pro derivaci složené funkce a s opětovným dosazením (2).

Členy na řádku (44) odpovídají chybě vnesené z předchozího kroku, tedy v tomto případě ε_1. Členy na řádku (45) představují lokální chybu ε_2 vzniklou v rámci druhého kroku. Zbývající členy na řádku (46) souvisejí s šířením chyby z předchozího výpočtu. Místo přesné derivace řešení v čase t_1, určené jako $f(u(t_1), t_1)$, pracujeme s hodnotou $f(u_1, t_1)$. Díky tomu, že funkce f je lipschitzovská s konstantou L (viz (13)), můžeme chybu při určení derivace $\dot{u}(t_1)$ odhadnout jako

$$|f(u(t_1), t_1) - f(u_1, t_1)| \leq L\,|u(t_1) - u_1| = L|\varepsilon_1| \tag{47}$$

V rovnici (43)–(46) je vyjádřena tzv. celková (tj. globální) chyba po druhém kroku, $\varepsilon_2^g = u(t_2) - u_2$, kterou nyní můžeme odhadnout takto:

$$|\varepsilon_2^g| \leq |\varepsilon_1| + |\varepsilon_2| + Lh_2|\varepsilon_1| \leq (1 + Lh_2)Ch_1^2 + Ch_2^2 \tag{48}$$

Ve třetím kroku můžeme postupovat podobně, ale za chybu na počátku kroku musíme dosadit globální chybu ε_2^g. Dostaneme tak odhad

$$\varepsilon_3^g \leq (1 + Lh_3)|\varepsilon_2^g| + |\varepsilon_3| \leq (1 + Lh_3)(1 + Lh_2)Ch_1^2 + (1 + Lh_3)Ch_2^2 + Ch_3^2 \tag{49}$$

Jestliže jsou všechny časové kroky stejné ($h_i = h$, $i = 1, 2 \ldots N$), dostaneme po n-tém kroku odhad

$$|\varepsilon_n^g| \leq Ch^2 \sum_{i=0}^{n-1} (1 + Lh)^i = Ch^2 \frac{(1 + Lh)^n - 1}{(1 + Lh) - 1} = \frac{Ch}{L}[(1 + Lh)^n - 1] \tag{50}$$

Připomeňte si, že s rostoucím k posloupnost $(1+1/k)^k$ konverguje k e, a to monotónně zdola. Výraz v hranaté závorce na pravé straně (50) lze tedy shora odhadnout jako $\mathrm{e}^{Lhn} - 1$, kde součin hn odpovídá času t_n po n krocích. Výsledný odhad globální chyby můžeme zapsat jako

$$|\varepsilon_n^g| \leq \frac{Ch}{L}\left(\mathrm{e}^{Lt_n} - 1\right) \tag{51}$$

Pro pevně zvolený čas tedy chyba nepřekročí jistý násobek délky kroku. Mluvíme o konvergenci řádu 1. Při desetinásobném zkrácení kroku lze očekávat, že chyba se sníží zhruba na desetinu. To ale platí jen asymptoticky, při dostatečně krátkém kroku.

Stabilita

Podle odhadu (51) sice při zkracování kroku h řešení pro pevně zvolený časový okamžik konverguje k přesné hodnotě, ale jestliže naopak zvolíme pevnou délku kroku h a zvyšujeme počet kroků (tedy i čas t_n), odhadnutá chyba rychle narůstá, pro dlouhé časy exponenciálně. Ve skutečnosti tomu tak být nemusí. Při odvození odhadu (51) jsme vycházeli z "nejpesimističtějšího scénáře", abychom postihli co nejširší třídu úloh. O funkci f určující charakter rovnice jsme předpokládali pouze, že je lipschitzovská.

V řadě prakticky významných úloh lze předpokládat, že f je nerostoucí funkcí proměnné u, což má pozitivní důsledky pro šíření chyby. Příkladem je rovnice popisující Kelvinův reologický model, tj. paralelně zapojenou pružinu a viskózní tlumič. Pomocí proměnných s fyzikálním významem se tato rovnice zapíše jako

$$\eta\dot{\varepsilon}(t) + E\varepsilon(t) = \sigma(t) \tag{52}$$

kde ε je deformace, σ je napětí, η je viskozita tlumiče a E je tuhost pružiny. Rovnici lze přepsat do tvaru

$$\dot{\varepsilon}(t) = \frac{\sigma(t)}{\eta} - \frac{E}{\eta}\varepsilon(t) \tag{53}$$

a pokud přeznačíme ε jako u, bude funkce f dána předpisem

$$f(u,t) = \frac{\sigma(t)}{\eta} - \frac{E}{\eta}u \tag{54}$$

Jelikož E a η jsou kladné konstanty, je f skutečně nerostoucí funkcí proměnné u. V tomto případě jde dokonce o funkci lineární vzhledem k u, protože jsme použili lineární pružinu a tlumič.

Pokud je f klesající funkce proměnné u, mají příspěvky k chybě v řádcích (44) a (46) opačná znaménka a nemusí docházet k šíření lokální chyby. Díky tomu se pak dá zlepšit výsledný odhad globální chyby. Stále předpokládáme, že f je lipschitzovská vzhledem k u s konstantou L. Je-li navíc nerostoucí vzhledem k u, lze místo (13) napsat

$$0 \geq \quad f(u_b) - f(u_a) \quad \geq -L(u_b - u_a) \qquad \text{pro } u_a < u_b \tag{55}$$

$$0 \leq \quad f(u_b) - f(u_a) \quad \leq -L(u_b - u_a) \qquad \text{pro } u_a > u_b \tag{56}$$

V obou případech (tedy nezávisle na znaménku $u_b - u_a$) je zaručeno, že

$$|u_b - u_a + h[f(u_b) - f(u_a)]| \leq |u_b - u_a| \max(1, Lh - 1) \tag{57}$$

Pokud platí

$$Lh \leq 2 \tag{58}$$

pak lze (57) zjednodušit na

$$|u_b - u_a + h[f(u_b) - f(u_a)]| \leq |u_b - u_a| \tag{59}$$

Tuto nerovnost můžeme využít při odhadu propagované lokální chyby, která je reprezentována součtem řádků (44) a (46). Stačí dosadit $u(t_1)$ za u_b a u_1 za u_a. Za předpokladu, že $Lh_2 \leq 2$, pak můžeme zaručit, že součet řádků (44) a (46) nebude v absolutní hodnotě větší než samotný řádek (44) a chyba z předchozího výpočtu se tedy nebude zvětšovat. Samozřejmě k ní ale opět přibude lokální chyba z dalšího kroku, reprezentovaná řádkem (45). Celkově tedy můžeme sestrojit vylepšený odhad chyb

$$|\varepsilon_2^g| \leq |\varepsilon_1| + |\varepsilon_2| \leq Ch_1^2 + Ch_2^2 \tag{60}$$

$$|\varepsilon_3^g| \leq |\varepsilon_2^g| + |\varepsilon_3| \leq Ch_1^2 + Ch_2^2 + Ch_3^2 \tag{61}$$

a obecně

$$|\varepsilon_n^g| \leq C \sum_{i=1}^{n} h_i^2 \tag{62}$$

Tyto odhady jsou platné za předpokladu, že všechny časové kroky splňují podmínku $Lh_n \leq 2$. Pokud jsou všechny kroky stejně velké a rovné h, bude

$$|\varepsilon_n^g| \leq Cnh^2 = Ct_n h \tag{63}$$

Za uvedených předpokladů tedy pro pevně zvolený krok a rostoucí čas chyba narůstá nejvýš lineárně (nikoli exponenciálně). Podmínkou je, aby byla splněna nerovnost (58), což znamená, že časový krok nesmí překročit kritickou hodnotu

$$h_{crit} = \frac{2}{L} \tag{64}$$

Při překročení kritické hodnoty časového kroku může dojít k exponenciálnímu nárůstu chyby v čase. V takovém případě mluvíme o ztrátě numerické stability. Eulerova dopředná metoda je pouze podmíněně stabilní, protože numerická stabilita je zajištěna jen pro dostatečně krátké kroky.

1.3 Eulerova zpětná (implicitní) metoda

Ukazuje se, že stabilitu numerického řešení lze zlepšit užitím zpětné varianty Eulerovy metody. Při tomto postupu se pravá strana rovnice (2) nahradí konstantní hodnotou určenou na konci kroku, což vede ke vztahu

$$u_n = u_{n-1} + h_n f(u_n, t_n), \qquad n = 1, 2, \ldots N \tag{65}$$

Jelikož hodnota u_n se vyskytuje i na pravé straně, nedá se do vzorce (65) jednoduše dosadit, ale je třeba jej chápat jako rovnici pro výpočet neznámé u_n. Proto mluvíme o implicitní metodě. Řešení je snadné, pokud je funkce f lineární vzhledem k proměnné u. Jinak je třeba u_n získat iterativním řešením, např. Newtonovou metodou, která ale vyžaduje znalost derivace $\partial f/\partial u$.

Pokud má funkce f lineární tvar

$$f(u, t) = a(t)u + b(t) \tag{66}$$

pak lze (65) přepsat jako

$$u_n = u_{n-1} + h_n \left[a(t_n)u_n + b(t_n)\right], \qquad n = 1, 2, \ldots N \tag{67}$$

a snadno odvodit explicitní výraz pro hodnotu řešení na konci kroku:

$$u_n = \frac{u_{n-1} + h_n b(t_n)}{1 - h_n a(t_n)} \tag{68}$$

Funkce f ve tvaru (66) je nerostoucí funkcí proměnné u, pokud je $a(t) \leq 0$ pro každé t. V tom případě je jmenovatel zlomku na pravé straně (68) větší nebo roven jedné a nedochází k šíření lokální chyby. Pozoruhodné je, že to platí nezávisle na délce kroku h_n. Říkáme proto, že Eulerova zpětná metoda je nepodmíněně stabilní. Řád konvergence je však i pro tuto metodu pouze 1.

1.4 Další metody

Eulerovu dopřednou i zpětnou metodu je možné chápat jako různé aproximace integrálu, který definuje přírůstek hodnoty funkce u podle (34). Aproximace

$$\int_{t_{n-1}}^{t_n} f(u(t), t)\, \mathrm{d}t \approx (t_n - t_{n-1})\, f(u_{n-1}, t_{n-1}) \tag{69}$$

založená na jednobodové integraci pomocí hodnoty integrandu na levém okraji vede k dopředné metodě, aproximace

$$\int_{t_{n-1}}^{t_n} f(u(t),t)\,\mathrm{d}t \approx (t_n - t_{n-1})\, f(u_n, t_n) \tag{70}$$

pak vede ke zpětné metodě. Řadu podobných metod je možno konstruovat podle vylepšených integračních schémat. Mezi **explicitní** metody patří například modifikovaná Eulerova metoda

$$u_n = u_{n-1} + h_n f\left(u_{n-1} + h_n f(u_{n-1}, t_{n-1})/2, t_{n-1} + h_n/2\right) \tag{71}$$

Heunova metoda

$$u_n = u_{n-1} + \frac{h_n}{2}\left[f(u_{n-1}, t_{n-1}) + f\left(u_{n-1} + h_n f(u_{n-1}, t_{n-1}), t_n\right)\right] \tag{72}$$

nebo Rungeova-Kuttaova metoda 4. řádu

$$u_n = u_{n-1} + \frac{h_n}{6}\left(k_1 + 2k_2 + 2k_3 + k_4\right) \tag{73}$$

kde

$$\begin{aligned} k_1 &= f(u_{n-1}, t_{n-1}) && (74)\\ k_2 &= f(u_{n-1} + h_n k_1/2, t_{n-1} + h_n/2) && (75)\\ k_3 &= f(u_{n-1} + h_n k_2/2, t_{n-1} + h_n/2) && (76)\\ k_4 &= f(u_{n-1} + h_n k_3, t_n) && (77) \end{aligned}$$

Implicitní metody mohou být založeny např. na zobecněném pravidle středního bodu (generalized midpoint rule = GMR)

$$u_n = u_{n-1} + h_n f\left((1-\alpha)u_{n-1} + \alpha u_n, t_{n-1} + \alpha h_n\right) \tag{78}$$

nebo na zobecněném lichoběžníkovém pravidle (generalized trapezoidal rule = GTR)

$$u_n = u_{n-1} + (1-\alpha)h_n f(u_{n-1}, t_{n-1}) + \alpha h_n f(u_n, t_n) \tag{79}$$

kde α je parametr mezi 0 a 1. Pro $\alpha = 0$ dostáváme dopřednou Eulerovu metodu a pro $\alpha = 1$ zpětnou Eulerovu metodu. Pro $\alpha = 1/2$ jde v případě rovnice (79) o standardní lichoběžníkové pravidlo a v případě rovnice (78) o standardní pravidlo středního bodu.

Jestliže je funkce f lineární vzhledem k proměnné u a má tvar (66), jsou rovnice (78)–(79) lineární a lze je vyřešit v uzavřeném tvaru. Například pro zobecněné pravidlo středního bodu (78) dostaneme

$$u_n = \frac{[1 + (1-\alpha)h_n a_{n-1+\alpha}]u_{n-1} + h_n b_{n-1+\alpha}}{1 - \alpha h_n a_{n-1+\alpha}} \tag{80}$$

kde jsme pro jednoduchost zápisu použili označení $a_{n-1+\alpha} = a(t_{n-1} + \alpha h_n)$ a $b_{n-1+\alpha} = b(t_{n-1} + \alpha h_n)$. Numerická stabilita je zaručena, pokud koeficient násobící u_{n-1} nepřekročí v absolutní hodnotě jedničku, tedy pokud

$$-1 \le \frac{1 + (1-\alpha)h_n a_{n-1+\alpha}}{1 - \alpha h_n a_{n-1+\alpha}} \le 1 \tag{81}$$

Pro $a(t) \leq 0$ to vede na podmínku

$$(2\alpha - 1)h_n a_{n-1+\alpha} \leq 2 \tag{82}$$

která je splněna pro $\alpha \geq 1/2$ při libovolně velkém kroku h_n a metoda je nepodmíněně stabilní. Pro $\alpha < 1/2$ pak dostáváme podmíněně stabilní metodu, přičemž kritický časový krok

$$h_{crit} = -\frac{2}{(1-2\alpha)a_{n-1+\alpha}} \tag{83}$$

se pro klesající α postupně zmenšuje až k minimální hodnotě $h_{crit} = -2/a_{n-1}$ pro $\alpha = 0$, což odpovídá dopředné Eulerově metodě, viz (64).

Pokud je součinitel násobící u_{n-1} záporný, lze očekávat oscilace v řešení. Tento případ nastává při délce kroku větší než

$$h_{osc} = -\frac{1}{(1-\alpha)a_{n-1+\alpha}} \tag{84}$$

Podobně pro zobecněné lichoběžníkové pravidlo a lineární funkci f ve tvaru (66) můžeme řešení rovnice (79) napsat jako

$$u_n = \frac{[1 + (1-\alpha)h_n a_{n-1}]u_{n-1} + (1-\alpha)h_n b_{n-1} + \alpha h_n b_n}{1 - \alpha h_n a_n} \tag{85}$$

kde $a_{n-1} = a(t_{n-1})$, $a_n = a(t_n)$, $b_{n-1} = b(t_{n-1})$ a $b_n = b(t_n)$. Za předpokladu $a(t) \leq 0$ lze podmínku stability

$$-1 \leq \frac{1 + (1-\alpha)h_n a_{n-1}}{1 - \alpha h_n a_n} \leq 1 \tag{86}$$

upravit na tvar

$$[\alpha a_n - (1-\alpha)a_{n-1}]h_n \leq 2 \tag{87}$$

Pokud je $a(t) = a = \text{const.}$, tedy pokud jde o lineární rovnici s konstantními koeficienty, jsou závěry podobné jako pro zobecněné pravidlo středního bodu: Metoda je nepodmíněně stabilní pro $\alpha \geq 1/2$ a podmíněně stabilní pro $\alpha < 1/2$, s kritickým krokem

$$h_{crit} = -\frac{2}{(1-2\alpha)a} \tag{88}$$

V případě proměnného koeficientu $a(t)$ lze splnění podmínky (87) zcela obecně zajistit pouze pro

$$h_n \leq -\frac{2}{(1-\alpha)a_{n-1}} \tag{89}$$

V tomto smyslu by tedy metoda byla jen podmíněně stabilní pro všechna $\alpha < 1$. Pro hodnoty $\alpha \geq 1/2$ ale k narušení stability může dojít jen při prudkém poklesu absolutní hodnoty funkce a v rámci daného kroku, který se však těžko může opakovat v celé řadě na sebe navazujících kroků. Pokud tedy metodu použijeme s časovým krokem nesplňujícím podmínku (89), může chyba v některých krocích narůst, ale nelze očekávat její systematický dramatický růst. Mohli bychom také pro zvolený krok h_n nesplňující podmínku (89) nejprve zkusmo spočítat $a_n = a(t_{n-1} + h_n)$ a zkontrolovat podmínku (87). Při jejím splnění lze krok bez obav provést, při nesplnění je lépe jej zkrátit a provést novou kontrolu podmínky (89).

Podrobnější analýza by ukázala, že obavy ze ztráty stability implicitní metody založené na zobecněném lichoběžníkovém pravidle jsou pro $\alpha \geq 1/2$ při konstantní délce kroku zbytečné. Lze to ilustrovat na jednoduchém příkladu. Zkoumejme například, jak chyba obsažená v aproximaci přesné hodnoty $u(t_3)$ přibližnou u_3 přispěje k chybě v aproximaci přesné hodnoty $u(t_6)$ přibližnou u_6. Při opakované aplikaci vzorce (85) bude původní chyba postupně násobena třemi faktory, z nichž každý má tvar zlomku z podmínky (86), přičemž n nabývá hodnot 4, 5 a 6. Celkově tedy bude chyba násobena faktorem

$$A_{3,6} = \frac{1+(1-\alpha)ha_3}{1-\alpha ha_4} \times \frac{1+(1-\alpha)ha_4}{1-\alpha ha_5} \times \frac{1+(1-\alpha)ha_5}{1-\alpha ha_6} \tag{90}$$

Pro $a(t) \leq 0$ a $\alpha \geq 1/2$ platí

$$\left|\frac{1+(1-\alpha)ha_4}{1-\alpha ha_4}\right| \leq 1, \qquad \left|\frac{1+(1-\alpha)ha_5}{1-\alpha ha_5}\right| \leq 1 \tag{91}$$

takže lze napsat odhad

$$|A_{3,6}| \leq \frac{|1+(1-\alpha)ha_3|}{1-\alpha ha_6} \leq \max(1, (1-\alpha)hL - 1) \tag{92}$$

Obdobně lze postupovat pro libovolný počet kroků, takže každá nově vzniklá lokální chyba nemůže neomezeně růst, ale může se "šířením" zvýšit maximálně na jistý násobek. Proto exponenciální růst globální chyby nehrozí.

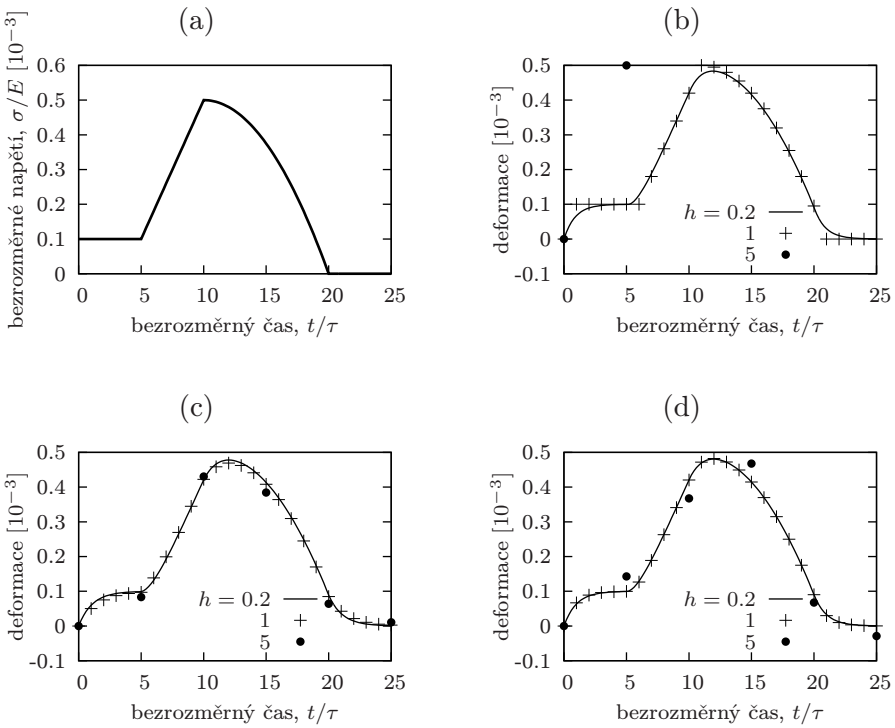
1.5 Příklad

Řešení Kelvinova modelu pomocí implicitní metody založené na zobecněném lichoběžníkovém pravidle

For illustration, the graphs in Fig. 2b-d show the numerical solutions of the strain history induced in the Kelvin model (see equation (53)) by the prescribed stress history specified in Fig. 2a. The stress scale is normalized by the spring stiffness E and the time scale by the retardation time $\tau = \eta/E$. The stress jumps to $10^{-4}E$ at time 0 and remains constant until time 5τ. Then it increases linearly to value $5 \times 10^{-4}E$ attained at time 10τ, and subsequently decreases quadratically and reaches zero at time 20τ. From that moment on, the stress remains at zero level.

The strain histories numerically computed by three special versions of the generalized trapezoidal rule (79) are presented in Fig. 2b-d. For each integration scheme, three different time steps (in this example kept constant during the entire solution) have been used. For the shortest time step, $h = 0.2\tau$, the results (plotted as solid curves) are almost the same for all the three methods, and are visually undistinguishable from the exact solution. Some differences can be observed for larger time steps. The forward Euler method (Fig. 2b) is inaccurate for $h = \tau$, and becomes unstable for $h = 5\tau$. Note that the Lipschitz constant is $L = E/\eta = 1/\tau$ and the critical time step according to (64) is $h_{\text{crit}} = 2/L = 2\tau$. The backward Euler method (Fig. 2c) is somewhat more accurate for $h = \tau$ and the error remains reasonable even for $h = 5\tau$. The standard trapezoidal rule leads to a good accuracy for $h = \tau$ (Fig. 2d), but large deviations of an oscillatory character appear for $h = 5\tau$.

The accuracy of the methods can be compared quantitatively if a specific measure of the error is defined and evaluated. In this example, we use the root mean square of the differences

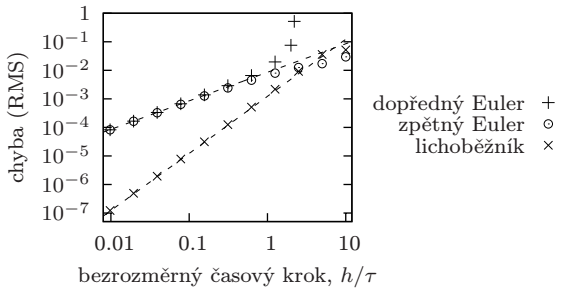


Obrázek 2: (a) Předepsaný vývoj napětí působícího na Kelvinův model; (b-d) vývoj deformace vypočtený (b) dopřednou Eulerovou metodou, (c) zpětnou Eulerovou metodou, (d) standardním lichoběžníkovým pravidlem

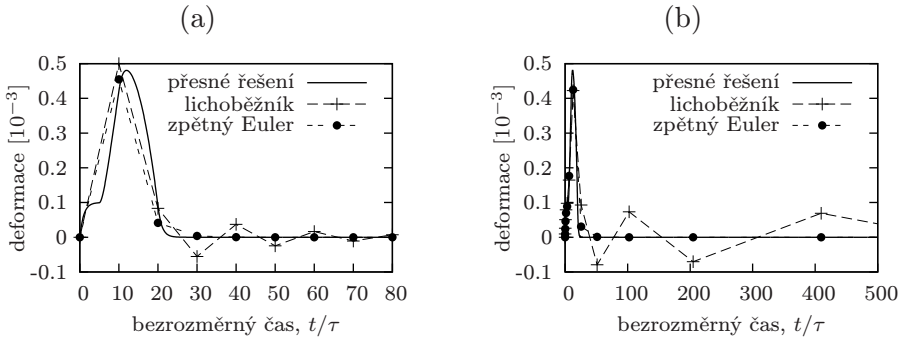
between the numerical solution and the exact one at all times t_k between 0 and 30τ. For a piecewise polynomial stress history, equation (53) admits an analytical solution, and so the error of the numerical scheme can be evaluated precisely. The dependence of the error on the step size is shown in logarithmic scale in the convergence diagram in Fig. 3. Individual points correspond to errors of the numerical schemes for different time steps, ranging from 0.01τ to 10τ. The dashed straight lines do not directly connect the points; they just indicate slopes 1:1 and 2:1 and help to identify the asymptotic convergence rates. As the step size tends to zero, the error of the forward as well as backward Euler methods is proportional to the step size, which means that in the logarithmic plot the points lie on a straight line of slope 1. For the standard trapezoidal rule, the error is proportional to the square of the step size and the points lie on a straight line of slope 2. This is in agreement with the theoretical analysis of the numerical scheme,[2] according to which $\alpha = 0.5$, corresponding to the standard trapezoidal rule (STR), is the only value that leads to a quadratic convergence rate, and for all the other values the convergence rate is linear.

The asymptotic convergence rate determines the error evolution as the step size tends to zero, and thus is related to accuracy for very short time steps. From this point of view,

[2]The numerical approximation of the integral in (34) based on the STR is exact for a linear function, and so the error of integration from t_k to $t_k + h$ is dominated by the quadratic part of the integrand and is proportional to h^3. The total number of time steps over an interval of fixed length is inversely proportional to h, and so the cumulative error is proportional to h^2. For all other versions of the generalized trapezoidal rule, with $\alpha \neq 0.5$, the integration is exact for a constant function only, and the error is due to the linear part of the integrand, thus being proportional to h^2 in one time step and to h after accumulation over a fixed interval.



Obrázek 3: Konvergenční diagram pro různé verze zobecněného lichoběžníkového pravidla (závislost chyby na délce kroku)



Obrázek 4: Porovnání vývoje deformace vypočteného pomocí standardního lichoběžníkového pravidla a zpětné Eulerovy metody s využitím (a) konstantní délky kroku 10τ, (b) počáteční délky kroku 0.1τ zvyšované v každém kroku na dvojnásobek

the STR is clearly superior to the forward or backward Euler methods. However, it is also interesting to look at the accuracy for medium or even large step sizes. As seen in Fig. 3, the error of the forward Euler method blows up for steps larger than $t_{\text{crit}} = 2\tau$, due to the loss of numerical stability. Already for step sizes below t_{crit} but comparable to it, the error is substantially larger than for the other methods; see also the strain values for $h = \tau$ in Fig. 2b. The STR gives the best accuracy for step sizes below 2.5τ, but for larger steps the backward Euler method seems to be superior. This can be confirmed by a computation with step size 10τ continued over a longer time interval (still using the prescribed stress history from Fig. 2a, with stress after time 20τ identically equal to zero). Of course, one cannot expect a high accuracy with such a long step, but the numerical results should at least reflect the main features of the solution, in particular they should quickly approach zero after time 20τ. This is indeed the case for the backward Euler scheme but not for the STR. As shown in Fig. 4a, the strain values computed by the STR oscillate even after complete removal of the stress, and the amplitude of oscillations decreases only slowly. The polution of the results by such oscillations becomes especially strong if the step size is progressively increased. Fig. 4b shows the strain values obtained when the initial step size $h_1 = 0.1\tau$ is doubled in each subsequent step. For the STR, a large error is still observed at times one or two orders of magnitude larger than the duration of the loading impulse. Again, the backward Euler method gives acceptable results.