Dokładność i wiarygodność obliczeń numerycznych

Wstęp

Wynik obliczeń numerycznych ma na ogół charakter *przybliżony*. Nawet w przypadku poprawnie sformułowanego modelu matematycznego wartość wyznaczona numerycznie może różnić się od wartości dokładnej wskutek niepewności danych wejściowych, skończonej precyzji arytmetyki komputerowej, aproksymacji właściwych danej metodzie numerycznej lub przedwczesnego zakończenia procesu iteracyjnego. Ponadto rozpatrywane zadanie może wykazywać znaczną wrażliwość na niewielkie zaburzenia danych, wskutek czego nawet zastosowanie algorytmu stabilnego numerycznie nie gwarantuje uzyskania wyniku o dużej dokładności [2, 3].

W obliczeniach z zakresu elektrotechniki wymienione zjawiska występują między innymi w następujących sytuacjach:

- pomiary napięcia i prądu są obarczone niepewnością, która propaguje się na wyznaczaną na ich podstawie wartość mocy;
- parametry elementów, takie jak rezystancja, indukcyjność i pojemność, są określone z ograniczoną dokładnością wynikającą między innymi z tolerancji wykonania;
- symulacja stanów przejściowych wymaga dyskretyzacji czasu, co prowadzi do błędu zależnego od wartości kroku czasowego;
- rozwiązanie rozbudowanego układu równań węzłowych może być wrażliwe na niewielkie zmiany przewodności gałęzi lub parametrów źródeł;
- odejmowanie napięć o zbliżonych wartościach może prowadzić do istotnej utraty cyfr znaczących.

Celem niniejszego rozdziału jest wykształcenie umiejętności identyfikowania podstawowych źródeł błędów w obliczeniach numerycznych oraz rozróżniania następujących zagadnień:

1. Jaka jest dokładność danych wejściowych?
2. W jakim stopniu rozpatrywane zagadnienie matematyczne jest wrażliwe na zaburzenia danych wejściowych?
3. Jakie błędy wynikają z zastosowanego algorytmu oraz ze specyfiki arytmetyki komputerowej?
4. W jaki sposób można ocenić wiarygodność otrzymanego wyniku?

Pojęcia i metody przedstawione w niniejszym rozdziale stanowią podstawę analiz prowadzonych w dalszych częściach opracowania, poświęconych numerycznemu rozwiązywaniu układów równań, interpolacji, całkowaniu i różniczkowaniu numerycznemu oraz rozwiązywaniu równań różniczkowych.

1 Reprezentacja zmiennoprzecinkowa

W arytmetyce zmiennoprzecinkowej liczba jest reprezentowana w postaci

$$x = (-1)^s \, m \, \beta^e, \tag{1}$$

gdzie s określa znak liczby, m oznacza jej część znaczącą, β jest podstawą systemu liczbowego, natomiast e oznacza wykładnik. W powszechnie stosowanej arytmetyce binarnej przyjmuje się $\beta = 2$.

Ze względu na skończoną liczbę bitów przeznaczonych na zapis części znaczącej i wykładnika zbiór liczb reprezentowalnych jest skończony. W konsekwencji większość liczb rzeczywistych podlega zaokrągleniu do jednej z liczb dostępnych w danym formacie.

Przy założeniu zaokrąglania do najbliższej liczby reprezentowalnej oraz przy pominięciu przepełnienia i niedomiaru operację tę opisuje standardowy model

$$\mathrm{fl}(x) = x(1+\delta), \qquad |\delta| \lesssim u, \tag{2}$$

gdzie $\mathrm{fl}(x)$ oznacza reprezentację zmiennoprzecinkową liczby x, natomiast u jest jednostką zaokrąglenia. W arytmetyce binarnej z zaokrąglaniem do najbliższej liczby jednostka u jest równa połowie odległości pomiędzy liczbą 1 a następną większą liczbą reprezentowalną.

Odległość pomiędzy liczbą 1 a najmniejszą liczbą reprezentowalną większą od 1 określa się mianem *epsilona maszynowego*. W środowisku Scilab wartość ta jest dostępna jako `%eps`.

```
format("e",16);
mprintf("epsilon maszynowy = %.16e\n",%eps);
mprintf("(1 + %%eps) - 1 = %.16e\n",(1+%eps)-1);
mprintf("(1 + %%eps/2) - 1 = %.16e\n",(1+%eps/2)-1);
```

Odległość pomiędzy kolejnymi liczbami zmiennoprzecinkowymi nie jest stała, lecz zależy od ich wykładnika. W obszarze liczb o dużych modułach odległość ta jest większa niż w sąsiedztwie liczb o małych modułach. Z tego względu dodanie niewielkiego przyrostu do liczby o odpowiednio dużym module może nie spowodować zmiany jej reprezentacji.

Operacje zmiennoprzecinkowe nie zachowują wszystkich własności działań w zbiorze liczb rzeczywistych. W szczególności dodawanie zmiennoprzecinkowe nie musi być łączne:

$$(a+b)+c \neq a+(b+c). \tag{3}$$

Przyczyną tej właściwości jest zaokrąglanie wyników pośrednich. Jeżeli w danej kolejności działań dodawane są liczby o znacznie różniących się modułach, składnik o małym module może nie wpłynąć na wynik reprezentowany w skończonej precyzji.

Przykład: Dla

$$a = 10^{16}, \qquad b = -10^{16}, \qquad c = 1$$

wyrażenie $(a+b)+c$ przyjmuje wartość 1, podczas gdy w typowej arytmetyce podwójnej precyzji wyrażenie $a+(b+c)$ przyjmuje wartość 0. W drugim przypadku składnik 1 zostaje utracony na etapie obliczania sumy $b+c$.

Kolejność sumowania rozbudowanych zbiorów danych może zatem wpływać na wynik obliczeń. Przykładowo zjawisko to zachodzić może podczas wyznaczania średniej z długiego przebiegu prądu, obliczania energii na podstawie całki z mocy oraz sumowania strat mocy w wielu elementach modelu.

Ograniczeniu podlega zarówno precyzja reprezentacji, jak i zakres możliwych do zapisania wartości. Wynik o module przekraczającym największą liczbę reprezentowalną prowadzi do *przepełnienia* (ang. *overflow*). Z kolei wynik o module mniejszym od najmniejszej dodatniej

liczby normalnej może zostać przedstawiony jako liczba subnormalna lub zaokrąglony do zera, co wiąże się ze zjawiskiem *niedomiaru* (ang. *underflow*). Standardy arytmetyki zmiennoprzecinkowej przewidują ponadto wartości specjalne, w tym nieskończoności oraz NaN (ang. *Not a Number*).

W analizie obwodów szczególnej kontroli wymaga między innymi dzielenie przez liczbę bliską zeru. Zależność

$$Z = \frac{U}{I}$$

może prowadzić do wartości o bardzo dużym module, gdy natężenie prądu I jest niewielkie, natomiast dla $I = 0$ może skutkować wartością nieskończoną. Taki rezultat nie musi świadczyć o błędzie implementacji; może wskazywać na szczególny przypadek modelu matematycznego, wymagający odrębnej interpretacji fizycznej.

2 Źródła błędów i niepewności

W analizie wiarygodności obliczeń numerycznych należy rozróżnić następujące źródła błędów i niepewności:

- **Błąd danych** wynika z rozbieżności pomiędzy wartościami wejściowymi a wartościami odpowiadającymi przyjętemu modelowi. Przykładowo jego źródłem mogą być niepewności pomiarów napięcia, prądu i rezystancji, tolerancje parametrów elementów oraz zaokrąglenia danych katalogowych.
- **Błąd zaokrąglenia** powstaje w procesie reprezentowania liczb i wykonywania działań w skończonej precyzji. Błędy tego rodzaju mogą kumulować się w wieloetapowych obliczeniach lub ulegać znacznemu wzmocnieniu wskutek niekorzystnej postaci numerycznej zastosowanego wzoru.
- **Błąd obcięcia** jest konsekwencją zastąpienia procesu granicznego jego skończonym przybliżeniem. Występuje między innymi w wyniku obcięcia szeregu Taylora oraz zastosowania niezerowego kroku w różniczkowaniu numerycznym lub numerycznym rozwiązywaniu równań różniczkowych.
- **Błąd iteracyjny** stanowi różnicę pomiędzy bieżącym przybliżeniem a granicą ciągu iteracyjnego. Jego wartość zależy między innymi od przyjętego kryterium zbieżności, tolerancji oraz maksymalnej liczby iteracji.
- **Błąd modelu** nie jest błędem numerycznym, lecz może mieć dominujący wpływ na zgodność wyniku z rzeczywistością. Powstaje wówczas, gdy model matematyczny pomija istotne zjawiska fizyczne, takie jak zależność rezystancji od temperatury, rezystancja przewodów, nieliniowość rdzenia magnetycznego czy pojemności i indukcyjności pasożytnicze.

Należy podkreślić, że wysoka dokładność rozwiązania numerycznego nie kompensuje niedokładności modelu matematycznego. Wynik obliczony z małym błędem numerycznym może zatem niewłaściwie opisywać rzeczywisty układ.

2.1 Błąd bezwzględny i względny

Dla wartości dokładnej x i jej przybliżenia $\widehat{x}$ błąd bezwzględny wynosi

$$e_{\mathrm{abs}} = |\widehat{x} - x|, \tag{4}$$

a błąd względny

$$e_{\mathrm{rel}} = \frac{|\widehat{x} - x|}{|x|}, \qquad x \neq 0. \tag{5}$$

Błąd bezwzględny jest wyrażany w jednostce rozpatrywanej wielkości, natomiast błąd względny jest wielkością bezwymiarową, często przedstawianą w postaci procentowej.

Przykład. Jeżeli napięcie wzorcowe wynosi 10,000 V, a obliczona lub zmierzona wartość wynosi 9,990 V, to

$$e_{\text{abs}} = 0{,}010\,\text{V}, \qquad e_{\text{rel}} = 0{,}001 = 0{,}1\%.$$

Ten sam błąd bezwzględny, równy 0,010 V, przy napięciu wzorcowym 0,020 V odpowiada błędowi względnemu równemu 50%. Przykład ten wskazuje, że błąd bezwzględny rozpatrywany bez odniesienia do skali analizowanej wielkości nie stanowi wystarczającej miary jakości wyniku.

Jeżeli x jest równe zeru lub ma moduł bliski zeru, błąd względny jest niezdefiniowany albo traci użyteczność interpretacyjną. W takich przypadkach należy stosować błąd bezwzględny lub błąd znormalizowany względem charakterystycznej skali rozpatrywanego zagadnienia.

W przypadku nieznanej wartości dokładnej bezpośrednie wyznaczenie błędu nie jest możliwe. Ocenę jakości wyniku przeprowadza się wówczas na podstawie wielkości dostępnych obliczeniowo, takich jak residuum, estymator błędu, różnice pomiędzy wynikami otrzymanymi dla różnych wartości kroku lub rozwiązanie referencyjne wyznaczone metodą o potwierdzonej, wyższej dokładności.

3 Utrata cyfr znaczących przy odejmowaniu

Istotnym zjawiskiem występującym w obliczeniach numerycznych jest *katastrofalna utrata cyfr znaczących*. Zachodzi ona podczas odejmowania liczb o zbliżonych wartościach, obarczonych uprzednio błędami zaokrągleń.

Rozważmy obliczenie różnicy

$$\Delta U = U_1 - U_2,$$

gdzie U_1 i U_2 mają duże, zbliżone wartości, natomiast ich różnica jest niewielka. Jeżeli informacja opisująca tę różnicę została utracona na etapie reprezentacji operandów, operacja odejmowania nie może jej odtworzyć. Względny błąd otrzymanej różnicy może być wówczas znacznie większy od względnych błędów reprezentacji obu napięć.

Przykład pomiarowy. Podczas pomiaru prądu z wykorzystaniem bocznika wyznaczany jest niewielki spadek napięcia, na przykład 50 mV, mimo że oba wejścia wzmacniacza pomiarowego mogą znajdować się na potencjale rzędu setek woltów względem masy. W układzie fizycznym dokładność pomiaru ograniczają przede wszystkim współczynnik tłumienia składowej wspólnej wzmacniacza, dopuszczalny zakres napięć wejściowych oraz rozdzielczość przetwornika. Numerycznym odpowiednikiem tego problemu jest utrata informacji o małej różnicy wskutek obecności znacznie większej składowej wspólnej.

Uwaga o przykładach numerycznych. W celu jednoznacznego zilustrowania tego zjawiska w arytmetyce podwójnej precyzji stosuje się niekiedy wartości składowej wspólnej znacznie przekraczające zakres napięć występujących w rzeczywistych układach, na przykład 10^{12}–10^{15} V. Wartości te służą wyłącznie do przeprowadzenia testu numerycznego i nie stanowią realistycznych parametrów obwodu.

Wpływ utraty cyfr znaczących można w wielu przypadkach ograniczyć przez algebraiczne przekształcenie wzoru do postaci korzystniejszej numerycznie. Należy również eliminować zbędne tworzenie wartości pośrednich o dużych modułach, jeżeli wynik końcowy stanowi ich niewielką różnicę.

4 Wpływ niepewności danych na wynik

Jeżeli wielkość wynikowa y jest funkcją wielkości wejściowych,

$$y = f(x_1, x_2, \ldots, x_n),$$

to wpływ niewielkich zmian danych wejściowych można, w ramach przybliżenia pierwszego rzędu, opisać za pomocą różniczki zupełnej:

$$\Delta y \approx \sum_{i=1}^{n} \frac{\partial f}{\partial x_i} \Delta x_i. \tag{6}$$

Oszacowanie maksymalnego odchylenia, przy założeniu niezależnego doboru znaków zaburzeń, można uzyskać z nierówności

$$|\Delta y| \lesssim \sum_{i=1}^{n} \left| \frac{\partial f}{\partial x_i} \right| \, |\Delta x_i|. \tag{7}$$

Jeżeli natomiast $u(x_i)$ oznaczają niepewności standardowe wielkości wejściowych, które są wzajemnie nieskorelowane, liniowe prawo propagacji niepewności przyjmuje postać

$$u^2(y) \approx \sum_{i=1}^{n} \left(\frac{\partial f}{\partial x_i} \right)^2 u^2(x_i). \tag{8}$$

Podejście probabilistyczne należy odróżnić od oszacowania najgorszego przypadku, ponieważ metody te opierają się na odmiennych założeniach i dostarczają wielkości o różnej interpretacji.

Przykład: moc odbiornika rezystancyjnego

Dla

$$P = \frac{U^2}{R} \tag{9}$$

różniczka względna wynosi w przybliżeniu

$$\frac{\Delta P}{P} \approx 2\frac{\Delta U}{U} - \frac{\Delta R}{R}. \tag{10}$$

Jeżeli interesuje nas najgorszy możliwy moduł odchylenia, otrzymujemy

$$\frac{|\Delta P|}{P} \lesssim 2\frac{|\Delta U|}{U} + \frac{|\Delta R|}{R}. \tag{11}$$

Współczynnik 2 występujący przy względnej zmianie napięcia wskazuje, że przy porównywalnych względnych niepewnościach pomiar napięcia może wnosić większy udział do niepewności wyniku niż pomiar rezystancji. Zależność ta wynika z kwadratowej zależności mocy od napięcia.

Przykład liczbowy. Dla $U = 230\,\mathrm{V}$, $R = 46\,\Omega$, względnej niepewności napięcia 0,5% i rezystancji 1% liniowe oszacowanie najgorszego przypadku daje

$$\frac{|\Delta P|}{P} \approx 2 \cdot 0{,}5\% + 1\% = 2\%.$$

Otrzymane oszacowanie wskazuje, że niezależnie od precyzji arytmetyki wykorzystanej do obliczeń wiarygodność końcowej wartości mocy jest ograniczona przez niepewność danych wejściowych.

5 Uwarunkowanie zadania

Uwarunkowanie określa wrażliwość rozwiązania matematycznego na niewielkie zaburzenia danych wejściowych. Jest ono właściwością *zadania matematycznego*, niezależną od zastosowanego algorytmu i jego implementacji.

W zadaniu dobrze uwarunkowanym niewielkim względnym zaburzeniom danych odpowiadają niewielkie względne zmiany rozwiązania. W zadaniu źle uwarunkowanym nawet małe zaburzenie danych może prowadzić do znacznej zmiany rozwiązania.

Rozważmy wyznaczanie rezystancji ze wzoru

$$R = \frac{U}{I}.$$

Jeżeli I jest dostatecznie odległe od zera względem skali błędu pomiarowego, niewielkie względne błędy U i I prowadzą na ogół do niewielkiego względnego błędu R. Gdy jednak wartość prądu jest porównywalna z jego błędem bezwzględnym, niepewność pomiaru może zdominować wyznaczoną wartość rezystancji. Źródłem trudności jest wówczas wrażliwość zależności matematycznej w rozpatrywanym punkcie pracy, a nie sama operacja dzielenia wykonywana przez komputer.

5.1 Układy równań liniowych

Dla układu

$$Ax = b \tag{12}$$

miarą wrażliwości jest wskaźnik uwarunkowania

$$\kappa(A) = \|A\| \, \|A^{-1}\|. \tag{13}$$

Dla małych zaburzeń danych wskaźnik $\kappa(A)$ wyznacza górne oszacowanie ich możliwego względnego wzmocnienia w rozwiązaniu, przy czym dokładna postać oszacowania zależy od rodzaju zaburzanych danych i przyjętej normy. Jeżeli $\kappa(A)$ jest bliski 1, układ jest dobrze uwarunkowany w danej normie. Duża wartość $\kappa(A)$ wskazuje natomiast na możliwość znacznej wrażliwości rozwiązania na zaburzenia.

W analizie obwodów rolę macierzy A może pełnić macierz przewodności otrzymana metodą potencjałów węzłowych. Złe uwarunkowanie może wystąpić między innymi w modelach obejmujących przewodności o bardzo różnych rzędach wielkości lub w przypadku dwóch węzłów silnie połączonych ze sobą, a jednocześnie słabo związanych z węzłem odniesienia.

W celu rozwiązania układu nie należy jawnie wyznaczać macierzy odwrotnej A^{-1}. Zamiast operacji

$$x = A^{-1}b$$

zaleca się zastosowanie procedury bezpośrednio rozwiązującej układ, na przykład operatora `A\b` w środowisku Scilab. Takie postępowanie jest na ogół korzystniejsze pod względem dokładności numerycznej i kosztu obliczeniowego [1].

6 Stabilność algorytmu

Stabilność numeryczna jest właściwością algorytmu. W ujęciu analizy wstecznej algorytm stabilny prowadzi do wyniku, który można interpretować jako dokładne rozwiązanie zadania o nieznacznie zaburzonych danych. Oznacza to, że algorytm nie wprowadza nadmiernego wzmocnienia błędów wynikających z arytmetyki skończonej precyzji.

Uwarunkowanie zadania i stabilność algorytmu są pojęciami odrębnymi:

- zastosowanie niestabilnego algorytmu do zadania dobrze uwarunkowanego może prowadzić do wyniku obarczonego znacznym błędem;

- rozwiązanie zadania źle uwarunkowanego może być silnie wrażliwe na dane nawet w przypadku zastosowania algorytmu stabilnego;
- algorytm stabilny nie eliminuje niekorzystnego uwarunkowania zadania, lecz ogranicza dodatkowy błąd wynikający z procesu obliczeniowego.

Interpretacja inżynierska. Jeżeli układ pomiarowy wykazuje niewielką czułość na badaną wielkość, zwiększenie precyzji obliczeń nie usuwa ograniczeń wynikających z właściwości układu fizycznego. Analogicznie, w zadaniu silnie wrażliwym na dane zastosowanie stabilnego algorytmu ogranicza dodatkową utratę dokładności, lecz nie zmienia uwarunkowania modelu matematycznego.

7 Residuum, błąd rozwiązania i błąd wsteczny

Dla przybliżonego rozwiązania $\widehat{x}$ układu liniowego

$$Ax = b$$

residuum wynosi

$$r = b - A\widehat{x}. \tag{14}$$

Residuum określa stopień spełnienia równania przez obliczone przybliżenie $\widehat{x}$. Nie stanowi ono jednak bezpośredniej miary odległości pomiędzy $\widehat{x}$ a rozwiązaniem dokładnym x.

Jeżeli $e = x - \widehat{x}$ jest błędem rozwiązania, to

$$Ae = r, \qquad e = A^{-1}r. \tag{15}$$

Zależność ta wyjaśnia, dlaczego mała norma residuum nie musi implikować małej normy błędu rozwiązania. Jeżeli norma A^{-1} jest duża, niewielkie residuum może odpowiadać znacznemu błędowi rozwiązania.

Praktyczne względne residuum można obliczyć następująco:

```
x=A\b;
r=b-A*x;
eta=norm(r)/(norm(A)*norm(x)+norm(b));
mprintf("względne residuum = %.6e\n",eta);
```

Wielkość η stanowi znormalizowaną miarę residuum i może być interpretowana jako oszacowanie normowego błędu wstecznego: określa względny poziom zaburzenia danych, dla którego obliczony wektor x byłby rozwiązaniem dokładnym odpowiednio zmodyfikowanego układu.

Przykład obwodowy. W przypadku równań węzłowych $Gv = J$ mała norma residuum oznacza, że dla obliczonych napięć węzłowych równania bilansu prądów są spełnione z dużą dokładnością numeryczną. Jeżeli jednak macierz G jest źle uwarunkowana, otrzymane napięcia mogą nadal znacznie różnić się od rozwiązania odpowiadającego niezaburzonym danym.

8 Kryteria stopu w metodach iteracyjnych

W metodach iteracyjnych konieczne jest określenie kryterium zakończenia obliczeń. Przykładowe kryterium oparte na normie residuum ma postać

$$\|r_k\| \le \varepsilon_{\text{abs}} + \varepsilon_{\text{rel}}\|b\|. \tag{16}$$

W powyższej nierówności występują dwa składniki:

- tolerancja bezwzględna ε_{abs}, która określa dopuszczalny poziom residuum także w przypadku $\|b\|$ bliskiego zeru;
- składnik względny $\varepsilon_{\text{rel}}\|b\|$, który dostosowuje wymagany poziom dokładności do skali prawej strony układu.

Wyłączne monitorowanie różnicy kolejnych przybliżeń

$$\|x_{k+1} - x_k\|$$

może prowadzić do błędnych wniosków. Niewielkie zmiany iteratów mogą wynikać ze stagnacji procesu, a nie z osiągnięcia dostatecznie dokładnego rozwiązania. Z tego względu kryterium oparte na residuum należy łączyć z ograniczeniem maksymalnej liczby iteracji oraz, w uzasadnionych przypadkach, z kryterium zmiany rozwiązania.

9 Zbieżność, rząd metody i dobór kroku

Metodę nazywa się zbieżną, jeżeli wyznaczane przez nią przybliżenie dąży do rozwiązania dokładnego przy zmniejszaniu parametru dyskretyzacji lub zwiększaniu liczby iteracji, zgodnie z definicją właściwą danej klasie metod. Dla metody rzędu p błąd globalny ma asymptotyczną postać

$$E(h) = Ch^p + \mathcal{O}(h^{p+1}). \tag{17}$$

Po dwukrotnym zmniejszeniu kroku, w zakresie asymptotycznym, zachodzi zależność

$$\frac{E(h)}{E(h/2)} \approx 2^p. \tag{18}$$

Oznacza to, że błąd metody pierwszego rzędu zmniejsza się w przybliżeniu dwukrotnie, natomiast błąd metody drugiego rzędu — w przybliżeniu czterokrotnie.

Przykład: numeryczne różniczkowanie napięcia

Dla przebiegu sinusoidalnego

$$u(t) = U_m \sin(\omega t)$$

pochodna dokładna wynosi

$$\frac{du}{dt} = U_m \omega \cos(\omega t).$$

Można ją przybliżyć różnicą w przód

$$D_{\text{p}}(h) = \frac{u(t+h) - u(t)}{h} \tag{19}$$

lub różnicą centralną

$$D_{\text{c}}(h) = \frac{u(t+h) - u(t-h)}{2h}. \tag{20}$$

Formuła różnicy w przód charakteryzuje się błędem obcięcia rzędu $\mathcal{O}(h)$, natomiast formuła różnicy centralnej — rzędu $\mathcal{O}(h^2)$. W zakresie asymptotycznym dokładność różnicy centralnej wzrasta zatem szybciej wraz ze zmniejszaniem kroku.

9.1 Ograniczenia związane ze zmniejszaniem kroku

Zmniejszanie wartości h redukuje błąd obcięcia, lecz jednocześnie zwiększa względne znaczenie błędów zaokrągleń. W ilorazie różnicowym odejmowane są coraz bardziej zbliżone wartości $u(t+h)$ i $u(t)$, po czym ich niewielka różnica jest dzielona przez małą wartość h.

W rezultacie zależność błędu całkowitego od kroku ma często przebieg zbliżony do kształtu litery „U”:

- dla dużego h dominuje błąd obcięcia;
- dla pośredniego h występuje minimum błędu;
- dla bardzo małego h zaczynają dominować błędy zaokrągleń.

Dobór kroku powinien być zatem oparty na analizie zbieżności, na przykład na porównaniu wyników dla h, $h/2$ i $h/4$, nie zaś na arbitralnym przyjęciu najmniejszej wartości możliwej do reprezentowania.

9.2 Obserwowany rząd zbieżności

Jeżeli znany jest błąd $E(h)$, obserwowany rząd można oszacować ze wzoru

$$p_{\text{obs}} = \frac{\log(E(h)/E(h/2))}{\log 2}. \tag{21}$$

Uzyskanie wartości p_{obs} zbliżonych do teoretycznego rzędu $p = 2$ wskazuje, że obliczenia prawdopodobnie znajdują się w zakresie asymptotycznym, a implementacja wykazuje przewidywaną zależność błędu od kroku. Wyraźny spadek lub nieregularność wartości p_{obs} dla bardzo małych kroków może świadczyć o dominującym wpływie błędów zaokrągleń.

10 Koszt obliczeniowy

Dokładność stanowi tylko jedno z kryteriów oceny metody numerycznej. Analiza powinna obejmować również koszt czasowy i pamięciowy.

Złożoność obliczeniowa opisuje asymptotyczny wzrost liczby operacji wraz z rozmiarem zadania n. Przykładowo klasyczna eliminacja Gaussa zastosowana do macierzy gęstej wymaga rzędu

$$\mathcal{O}(n^3)$$

operacji oraz

$$\mathcal{O}(n^2)$$

pamięci.

W modelach rozległych sieci elektrycznych występują zazwyczaj macierze rzadkie, ponieważ każdy węzeł jest bezpośrednio połączony jedynie z niewielką częścią pozostałych węzłów. Przechowywanie takiej macierzy w formacie gęstym byłoby nieefektywne. Stosuje się zatem formaty rzadkie oraz algorytmy, których koszt zależy przede wszystkim od liczby elementów niezerowych i struktury ich rozmieszczenia.

Rzetelne porównanie metod powinno zatem uwzględniać jednocześnie:

- osiągnięty błąd lub residuum;
- liczbę iteracji, kroków albo wywołań funkcji;
- czas obliczeń;
- w większych zadaniach również zużycie pamięci.

Większa dokładność pojedynczego kroku nie implikuje mniejszego całkowitego czasu obliczeń. Metoda wyższego rzędu może wymagać większej liczby operacji w każdym kroku, a zarazem osiągać zadaną dokładność przy zastosowaniu znacznie większego kroku i mniejszej liczby kroków.

11 Minimalny protokół weryfikacji wyniku

Dla każdego zadania laboratoryjnego zaleca się stosowanie jednolitego protokołu weryfikacji:

1. **Zdefiniować dane wejściowe** wraz z ich skalą, jednostkami oraz, jeżeli jest to zasadne, niepewnością.
2. **Określić miarę jakości wyniku**: błąd bezwzględny, względny, residuum albo inną normę.
3. **Podać parametry algorytmu**: krok, tolerancję, maksymalną liczbę iteracji i zastosowaną metodę.
4. **Obliczyć residuum lub błąd** względem rozwiązania wzorcowego, jeżeli jest ono znane.
5. **Wykonać test czułości numerycznej**: zmienić krok, tolerancję, precyzję albo niewielkie zaburzenie danych.
6. **Ocenić uwarunkowanie**, jeżeli wymaga tego charakter zadania, oraz odnotować ostrzeżenia dotyczące osobliwości lub bliskości macierzy do osobliwości.
7. **Ocenić koszt obliczeń**, uwzględniając liczbę iteracji, czas wykonania oraz, w razie potrzeby, wykorzystanie pamięci.
8. **Sformułować wniosek fizyczny**, określający, czy dokładność numeryczna jest adekwatna do dokładności modelu i jakości danych pomiarowych.

Ostatni etap zapobiega nadinterpretacji dokładności numerycznej. Przykładowo raportowanie wartości mocy z rozdzielczością do mikrowata nie jest uzasadnione, jeżeli napięcie i rezystancja są określone z niepewnością względną rzędu jednego procenta.

12 Przykłady uzupełniające

Przykład 1. Rezystancja wyznaczana z dwóch pomiarów

Dla

$$R = \frac{U}{I}$$

przyjmijmy

$$U = 12{,}00\,\mathrm{V}, \qquad I = 2{,}00\,\mathrm{A}.$$

Wtedy $R = 6{,}00\,\Omega$. Jeżeli oba mierniki mają względną niepewność 1%, najgorszy przypadek daje w przybliżeniu

$$\frac{|\Delta R|}{R} \lesssim \frac{|\Delta U|}{U} + \frac{|\Delta I|}{I} = 2\%.$$

Zastosowanie w programie typu `double` nie prowadzi do zmniejszenia niepewności wynikającej z danych wejściowych. Precyzja arytmetyki może bowiem znacznie przewyższać dokładność informacji zawartej w danych pomiarowych.

Przykład 2. Energia z próbek mocy

Energię można aproksymować sumą

$$E \approx \Delta t \sum_{k=0}^{N-1} p_k.$$

W powyższej formule występują co najmniej trzy niezależne źródła błędu:

1. błędy pomiaru próbek p_k;
2. błąd obcięcia wynikający z zastąpienia całki sumą;
3. błąd zaokrąglenia podczas sumowania bardzo dużej liczby składników.

Jeżeli liczba próbek N jest bardzo duża, kolejność sumowania może wpływać na najmniej znaczące cyfry wyniku. W celu ograniczenia tego efektu można sumować składniki w kolejności rosnących modułów lub zastosować algorytm sumowania kompensowanego, na przykład algorytm Kahana.

Przykład 3. Równania węzłowe

Dla prostego modelu liniowego zapisujemy

$$Gv = J,$$

gdzie G jest macierzą przewodności, v wektorem napięć węzłowych, a J wektorem wymuszeń prądowych. Po rozwiązaniu układu należy sprawdzić

$$r = J - Gv.$$

Duża wartość $\|r\|$ oznacza, że otrzymany wynik nie spełnia z odpowiednią dokładnością równań wynikających z praw Kirchhoffa w przyjętym modelu. Jeżeli $\|r\|$ jest małe, lecz $\kappa(G)$ ma dużą wartość, konieczne jest dodatkowe zbadanie wrażliwości napięć na niewielkie zmiany parametrów.

Przykład 4. Dyskretyzacja równania obwodu RC

Dla rozładowania kondensatora przez rezystor zachodzi

$$\frac{du_C}{dt} = -\frac{1}{RC} u_C.$$

Najprostsza metoda Eulera daje

$$u_C^{k+1} = u_C^k - \frac{h}{RC} u_C^k.$$

Wielkość h oznacza krok czasowy. Zastosowanie zbyt dużego kroku prowadzi do znacznego błędu dyskretyzacji, a w przypadku niespełnienia warunków stabilności może również powodować jakościowo niepoprawne zachowanie rozwiązania numerycznego. Dla jawnej metody Eulera zastosowanej do rozpatrywanego równania warunkiem stabilności jest $0 < h < 2RC$; monotoniczny zanik bez zmiany znaku przybliżenia wymaga dodatkowo $0 < h \le RC$. Zmniejszenie kroku na ogół poprawia dokładność odwzorowania przebiegu, lecz zwiększa liczbę kroków i koszt obliczeń, co ilustruje typowy kompromis pomiędzy dokładnością a nakładem obliczeniowym.

Zadania

Zadania należy zrealizować w ramach jednego skryptu obliczeniowego. Każdy student stosuje w całym ćwiczeniu parametry z jednego wiersza tabeli. W zadaniach, w których wprost podano inne wartości, parametr zestawu zastępuje odpowiednią wartość nominalną. Dla każdego eksperymentu należy zarejestrować dane wejściowe, uzyskany wynik, przyjętą miarę błędu lub residuum oraz zwięzły wniosek. Wyniki zaleca się przedstawiać w postaci tabelarycznej.

Zestaw	U [V]	R [Ω]	δ_U [%]	δ_R [%]	t_0 [ms]	R_{RC} [kΩ]
1	230	46	0,50	1,00	3,70	1,00
2	220	44	0,55	0,90	3,50	1,10
3	240	48	0,45	1,10	3,90	0,90
4	110	22	0,60	1,20	2,50	1,20
5	120	24	0,40	0,80	2,80	0,80
6	200	50	0,70	1,00	3,20	1,30
7	210	42	0,35	1,30	3,40	0,70
8	225	45	0,65	0,70	3,60	1,40
9	235	47	0,50	1,40	3,80	0,60
10	250	50	0,30	0,60	4,00	1,50
11	180	36	0,75	1,10	3,10	0,95
12	150	30	0,55	0,85	2,90	1,05
13	260	52	0,40	1,25	4,10	0,85
14	100	20	0,80	1,50	2,20	1,25
15	270	54	0,25	0,75	4,30	1,15

We wszystkich zestawach dla obwodu RC przyjąć $C = 100\,\mu\text{F}$ i $U_0 = 10$ V; w eksperymencie różniczkowania zachować $U_m = 325$ V i $f = 50$ Hz, zmieniając tylko t_0.

1. **Epsilon maszynowy i niełączność dodawania.**

   (a) Wyświetlić wartość `%eps`, a następnie wyznaczyć ją eksperymentalnie. Rozpocząć od `delta=1` i dzielić wartość `delta` przez dwa tak długo, aż w arytmetyce komputera zostanie spełniona równość `1+delta==1`. Porównać ostatnią wartość zmieniającą wynik z `%eps`.

   (b) Dla
   $$a = 10^{16}, \qquad b = -10^{16}, \qquad c = 1$$
   obliczyć $(a + b) + c$ oraz $a + (b + c)$. Wyjaśnić, dlaczego wyniki są różne, mimo że dodawanie liczb rzeczywistych jest łączne.

   (c) Powtórzyć obliczenia dla kilku wartości $a = 10^k$. Określić najmniejsze k, dla którego oba sposoby sumowania przestają dawać ten sam wynik.

   (d) **Interpretacja elektrotechniczna:** wyjaśnić, dlaczego podobny problem może wystąpić przy sumowaniu bardzo dużej liczby małych przyrostów energii do dużej wartości skumulowanej albo przy sumowaniu próbek o bardzo różnych modułach.

   Na co zwrócić uwagę. Wniosku nie należy ograniczać do ogólnego stwierdzenia o niedokładności obliczeń komputerowych. Należy wykazać, że każdy wynik pośredni jest reprezentowany z wykorzystaniem skończonej liczby cyfr, a kolejność wykonywania działań wpływa na zakres informacji utraconej wskutek zaokrągleń.

2. **Utrata cyfr znaczących.**

   W układzie pomiarowym mała różnica dwóch sygnałów jest wyznaczana na tle dużej składowej wspólnej. Przyjąć wartość różnicową
   $$\Delta U = 0{,}1\,\text{V}$$

i kolejno

$$U_{\text{wsp}} = 10^3,\ 10^6,\ 10^9,\ 10^{12},\ 10^{15}\ \text{V}.$$

Podane wartości mają charakter testu numerycznego. Szczególnie największe z nich nie są realistycznymi napięciami układu elektrycznego; służą do ujawnienia ograniczeń arytmetyki zmiennoprzecinkowej.

(a) Dla każdej wartości utworzyć

$$U_1 = U_{\text{wsp}} + \Delta U, \qquad U_2 = U_{\text{wsp}},$$

a następnie obliczyć

$$\widehat{\Delta U} = U_1 - U_2.$$

(b) Wyznaczyć błąd bezwzględny i względny obliczonej różnicy.

(c) Dodatkowo obliczyć względny błąd reprezentacji samego U_1. Porównać go ze względnym błędem otrzymanej różnicy. Wyjaśnić, dlaczego małe względne błędy dwóch dużych liczb mogą dać duży względny błąd ich małej różnicy.

(d) Wskazać, od jakiej wartości składowej wspólnej wynik traci zauważalną liczbę cyfr znaczących.

(e) Odnieść wynik do pomiaru napięcia na boczniku, wejścia różnicowego wzmacniacza pomiarowego albo różnicy dwóch dużych mocy. Wyjaśnić, które ograniczenia w rzeczywistym torze pomiarowym są analogiczne do problemu numerycznego, a które wynikają z fizycznych parametrów aparatury.

3. **Wpływ niepewności pomiarów na obliczoną moc.**

Dla odbiornika rezystancyjnego przyjąć wartości nominalne

$$U = 230\,\text{V}, \qquad R = 46\,\Omega, \qquad P = \frac{U^2}{R}.$$

Niepewność względna pomiaru napięcia wynosi 0,5%, a rezystancji 1%.

(a) Obliczyć nominalną moc odbiornika.

(b) Wyznaczyć moc dla czterech kombinacji wartości granicznych:

$$U(1 \pm 0{,}005), \qquad R(1 \pm 0{,}01).$$

(c) Podać najmniejszą i największą moc oraz maksymalne względne odchylenie od wartości nominalnej.

(d) Porównać wynik z liniowym oszacowaniem

$$\frac{|\Delta P|}{P} \approx 2\frac{|\Delta U|}{U} + \frac{|\Delta R|}{R}.$$

Ocenić, który pomiar ma większy wpływ na niepewność obliczonej mocy.

(e) Powtórzyć analizę przy dwukrotnie dokładniejszym pomiarze napięcia, pozostawiając niepewność rezystancji bez zmian. Sprawdzić, o ile poprawia się oszacowanie mocy.

(f) Skomentować, czy zwiększenie liczby cyfr wyświetlanych przez program zwiększa rzeczywistą dokładność wyniku, jeżeli nie zmienia się jakość danych wejściowych.

4. **Uwarunkowanie i residuum modelu obwodu.**

   Rozpatrzyć dwa układy równań węzłowych:

$$G_1 = \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}, \qquad G_2 = \begin{bmatrix} 1{,}000001 & -1 \\ -1 & 1{,}000001 \end{bmatrix}, \qquad J = \begin{bmatrix} 1 \\ -1 \end{bmatrix}.$$

   Macierz G_2 odpowiada dwóm węzłom połączonym dużą przewodnością, które są połączone z węzłem odniesienia bardzo małymi przewodnościami. W takiej sytuacji niektóre kombinacje napięć mogą być słabo określone przez równania.

   (a) Dla obu macierzy obliczyć `cond(G)` i rozwiązanie `v=G\J`.

   (b) Zaburzyć wektor wymuszeń:

$$\widetilde{J} = J + 10^{-6} \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

   i ponownie rozwiązać oba układy.

   (c) Obliczyć względną zmianę danych

$$\frac{\|\widetilde{J} - J\|}{\|J\|}$$

   oraz względną zmianę rozwiązania

$$\frac{\|\widetilde{v} - v\|}{\|v\|}.$$

   (d) Porównać iloraz względnej zmiany rozwiązania i względnej zmiany danych z wartością `cond(G)`. Nie należy oczekiwać dokładnej równości — wskaźnik uwarunkowania opisuje najgorsze możliwe wzmocnienie w danej normie.

   (e) Dla $\widetilde{v}$ obliczyć residuum względem pierwotnego układu

$$r = J - G\widetilde{v}.$$

   Wyjaśnić, dlaczego w zadaniu źle uwarunkowanym małe residuum nie musi oznaczać małego błędu napięć.

   (f) Sprawdzić również residuum $\widetilde{J} - G\widetilde{v}$. Wyjaśnić, dlaczego powinno ono być bardzo małe, mimo że $\widetilde{v}$ może istotnie różnić się od v.

Interpretacja fizyczna. Warto rozróżnić dwa zjawiska. Pierwsze to poprawność rozwiązania numerycznego dla danych, które faktycznie przekazano do programu. Drugie to wrażliwość rozwiązania na to, czy te dane są dokładne. Układ może zostać rozwiązany z bardzo małym residuum, a mimo to niewielki błąd modelu przewodności lub źródeł może powodować dużą zmianę obliczonych napięć.

5. **Eksperyment krokowy dla sygnału napięciowego.**

   Napięcie sinusoidalne opisano funkcją

$$u(t) = 325 \sin(2\pi 50 t).$$

   Dla chwili

$$t_0 = 0{,}0037\,\mathrm{s}$$

   wyznaczyć pochodną dokładną

$$u'(t_0) = 325 \cdot 2\pi 50 \cos(2\pi 50 t_0),$$

   a następnie jej przybliżenia:

$$D_{\mathrm{p}}(h) = \frac{u(t_0 + h) - u(t_0)}{h}, \qquad D_{\mathrm{c}}(h) = \frac{u(t_0 + h) - u(t_0 - h)}{2h}.$$

(a) Wykonać obliczenia dla

$$h = 10^{-3},\ 5 \cdot 10^{-4},\ 2{,}5 \cdot 10^{-4},\ 1{,}25 \cdot 10^{-4}\ \text{s}.$$

(b) Dla obu wzorów obliczyć błąd bezwzględny. Wyniki zestawić w tabeli i przedstawić na wykresie logarytmicznym błędu w funkcji kroku.

(c) Dla kolejnych par kroków wyznaczyć obserwowany rząd

$$p_{\text{obs}} = \frac{\log(E(h)/E(h/2))}{\log 2}.$$

Porównać rząd różnicy w przód i różnicy centralnej.

(d) Rozszerzyć eksperyment o mniejsze kroki, np. aż do $10^{-12}\,\text{s}$, stosując skalę logarytmiczną. Sprawdzić, czy błąd maleje bez końca. Jeżeli zacznie rosnąć, wskazać moment przejścia od dominacji błędu obcięcia do dominacji błędu zaokrągleń.

(e) Wskazać, która metoda osiąga zadaną dokładność przy mniejszej liczbie wywołań funkcji. Skomentować zależność pomiędzy krokiem, błędem i kosztem.

(f) Zinterpretować $u'(t)$ fizycznie. Dla napięcia na kondensatorze zależność $i = C\,du/dt$ oznacza, że błąd numerycznego różniczkowania napięcia przekłada się bezpośrednio na błąd obliczonego prądu kondensatora.

6. **Błąd obcięcia w symulacji obwodu RC.**

Rozpatrzyć rozładowanie kondensatora opisane równaniem

$$\frac{du_C}{dt} = -\frac{1}{RC}u_C, \qquad u_C(0) = U_0.$$

Przyjąć

$$R = 1\,\text{k}\Omega, \qquad C = 100\,\mu\text{F}, \qquad U_0 = 10\,\text{V}.$$

Rozwiązanie dokładne wynosi

$$u_C(t) = U_0 \exp\!\left(-\frac{t}{RC}\right).$$

(a) Zaimplementować jawną metodę Eulera

$$u_C^{k+1} = u_C^k - \frac{h}{RC}u_C^k.$$

(b) Obliczyć przebieg do czasu $t = 5RC$ dla kilku kroków, np.

$$h = 0{,}5RC,\ 0{,}2RC,\ 0{,}1RC,\ 0{,}05RC.$$

(c) Dla każdego kroku obliczyć maksymalny błąd bezwzględny względem rozwiązania dokładnego.

(d) Sprawdzić, jak zmienia się błąd po dwukrotnym zmniejszeniu kroku i oszacować obserwowany rząd metody.

(e) Porównać dokładność z liczbą wykonanych kroków. Sformułować wniosek o kompromisie między kosztem a błędem obcięcia.

7. **Sumowanie energii i wpływ kolejności działań.**

Załóżmy, że moc chwilowa została zapisana jako duży zbiór próbek p_k, a energia jest liczona ze wzoru

$$E = \Delta t \sum_{k=1}^{N} p_k.$$

(a) Wygenerować zbiór zawierający bardzo dużo małych wartości oraz niewielką liczbę wartości o znacznie większym module.

(b) Obliczyć sumę w kolejności pierwotnej, po uporządkowaniu według rosnącego modułu i według malejącego modułu.

(c) Jeżeli środowisko na to pozwala, zaimplementować sumowanie kompensowane Kahana i porównać wyniki.

(d) Za rozwiązanie referencyjne przyjąć wynik uzyskany w wyższej precyzji albo dla konstrukcji danych, dla której suma dokładna jest znana analitycznie.

(e) Wyjaśnić, dlaczego matematycznie równoważne kolejności dodawania mogą dawać różne wyniki numeryczne i w jakich zastosowaniach elektrotechnicznych ma to znaczenie.

Podsumowanie ćwiczenia. Na zakończenie należy przygotować tabelę zawierającą dla każdego zadania:

- analizowane źródło błędu lub niepewności;
- zastosowaną miarę jakości wyniku;
- najważniejszy wynik liczbowy;
- obserwowaną zależność, np. wpływ kroku, uwarunkowania lub kolejności działań;
- jednozdaniowy wniosek inżynierski.

W podsumowaniu należy jednoznacznie rozróżnić:

1. błąd danych i niepewność pomiarową;
2. błąd zaokrąglenia;
3. błąd obcięcia;
4. błąd iteracyjny;
5. uwarunkowanie zadania;
6. stabilność algorytmu;
7. błąd modelu fizycznego.

Podstawowym celem analizy numerycznej nie jest uzyskanie możliwie dużej liczby cyfr po separatorze dziesiętnym, lecz określenie, *ile cyfr otrzymanego wyniku można uznać za wiarygodne oraz jakie przesłanki uzasadniają tę ocenę.*

Bibliografia

[1] Gene H. Golub i Charles F. Van Loan. *Matrix Computations*. 4 wyd. Baltimore: Johns Hopkins University Press, 2013.

[2] Nicholas J. Higham. *Accuracy and Stability of Numerical Algorithms*. 2 wyd. Philadelphia: SIAM, 2002.

[3] Alfio Quarteroni, Riccardo Sacco i Fausto Saleri. *Numerical Mathematics*. 2 wyd. Berlin: Springer, 2007.