Fundusze Europejskie dla Rozwoju Społecznego
Rzeczpospolita Polska
Dofinansowane przez Unię Europejską

Politechnika Świętokrzyska
Wydział Zarządzania i Modelowania Komputerowego

Kierunek studiów:
Inżynieria Danych

Paweł Stąpór

Materiały dydaktyczne do przedmiotu

METODY NUMERYCZNE

opracowane w ramach realizacji Projektu
„Dostosowanie kształcenia w Politechnice Świętokrzyskiej do potrzeb współczesnej gospodarki”
FERS.01.05-IP.08-0234/23

Kielce, 2025

Spis treści

1. Wstęp

Metody numeryczne to dziedzina matematyki stosowanej, która zajmuje się opracowywaniem i analizą algorytmów służących do rozwiązywania problemów matematycznych za pomocą obliczeń numerycznych a więc idealnych do implementacji w postaci programu komputerowego. W przeciwieństwie do metod analitycznych, które często dają dokładne rozwiązania w postaci wzorów, metody numeryczne najczęściej koncentrują się na uzyskiwaniu przybliżonych rozwiązań z określoną dokładnością. Jest to szczególnie przydatne, gdy dokładne rozwiązania są trudne lub niemożliwe do uzyskania analitycznie.

Główne zagadnienia podejmowane w tym przedmiocie obejmują:

- **Błędy w obliczeniach numerycznych:** Zrozumienie natury błędów (zaokrągleń, ucięcia), ich propagacji i wpływu na dokładność wyników.
- **Rozwiązywanie układów równań liniowych:** Metody bezpośrednie (Cramera, eliminacja Gaussa) i iteracyjne (Jacobiego, Gaussa-Seidela).
- **Rozwiązywanie równań nieliniowych:** Metody iteracyjne do znajdowania pierwiastków (bisekcji, iteracji prostej, Newtona-Raphsona).
- **Aproksymacja i interpolacja funkcji:** Tworzenie funkcji (np. wielomianów), które najlepiej pasują do danego zbioru punktów.
- **Numeryczne całkowanie i różniczkowanie:** Przybliżone metody obliczania całek oznaczonych i pochodnych funkcji.
- **Rozwiązywanie równań różniczkowych zwyczajnych problemów początkowy:** Metody Rungego-Kutty, Eulera.
- **Rozwiązywanie równań różniczkowych zwyczajnych problemów brzegowych:** Metoda różnic skończonych, Metoda elementów skończonych.

W praktycznym zastosowaniu metod numerycznych często wykorzystuje się specjalistyczne oprogramowanie. Do najczęściej wykorzystywanych zaliczyć można:

- **MATLAB/Octave:** Popularne środowiska do obliczeń numerycznych, oferujące szeroki zakres funkcji do implementacji i analizy algorytmów.
- **Python z bibliotekami NumPy, SciPy, Matplotlib:** Elastyczne i potężne narzędzie, które zdobywa coraz większą popularność w nauce i inżynierii.

Niniejsze materiały stanowią pomoc dydaktyczna dla studentów kierunku Inżynieria danych uczących się przedmiotu Metody numeryczne.

2. Błędy w obliczeniach numerycznych

Błędy w obliczeniach numerycznych są nieodłączną częścią pracy z komputerami i innymi systemami cyfrowymi. Wynikają one z ograniczeń w reprezentacji liczb oraz z natury algorytmów numerycznych, które często opierają się na przybliżeniach. Zrozumienie tych błędów i umiejętność ich szacowania jest kluczowe w metodach numerycznych.

2.1. Rodzaje błędów w obliczeniach numerycznych

Błędy danych wejściowych (początkowe)

Błędy wynikające z niedokładności danych, które są wprowadzane do obliczeń. Może to być spowodowane błędami pomiarowymi (np. z urządzeń fizycznych) lub koniecznością zaokrąglenia liczb niewymiernych czy nieskończonych do skończonej liczby cyfr.

Przykład 2.1. Wprowadzenie liczby π z ograniczoną precyzją (np. 3.14).

Błędy wynikające z uproszczeń przyjętego modelu

Przykład 2.2. Popularna postać równania matematycznego opisującego ruch wahadła, w którym sinus kąta wychylenia wahadła jest przybliżany wartością samego kąta wychylenia α, ponieważ $\sin(\alpha)\approx\alpha$ dla małych wartości kąta α wyrażonego w radianach, a więc zamiast rozwiązywać nieliniowe równanie

$$l\frac{\mathrm{d}^2\alpha}{\mathrm{d}t^2}+g\sin(\alpha)=0 \tag{2.1}$$

możemy rozwiązać jego prostszą liniową formę

$$l\frac{\mathrm{d}^2\alpha}{\mathrm{d}t^2}+g\alpha=0, \tag{2.2}$$

gdzie l oznacza długość wahadła, g przyspieszenie ziemskie a t czas.

Błędy reprezentacji (maszynowe)

Powstają, ponieważ komputery przechowują liczby w skończonej liczbie bitów (np. 32-bitowe lub 64-bitowe liczby zmiennoprzecinkowe). Wiele liczb rzeczywistych nie ma dokładnego odpowiednika w systemie binarnym, co wymusza zaokrąglanie.

Przykład: Liczba 0.1 w systemie dziesiętnym ma nieskończone rozwinięcie binarne, co prowadzi do błędu reprezentacji w systemie binarnym.

Błędy obcięcia (*truncation errors*)

Pojawiają się, gdy nieskończone procesy matematyczne (np. szeregi nieskończone, całki, pochodne) są przybliżane za pomocą skończonej liczby kroków lub wyrazów.

Przykład: Obliczanie wartości funkcji e^x za pomocą nieskończonego szeregu Maclaurina:

$$e^x = \sum_{n=0}^{\infty} \frac{1}{n!} x^n = 1 + x + \frac{1}{2!}x^2 + \frac{1}{3!}x^3 + \ldots \tag{2.3}$$

z wykorzystaniem skończonej liczby wyrazów

$$e^x \approx 1 + x + \frac{1}{2!}x^2 + \frac{1}{3!}x^3 + \ldots + \frac{1}{N!}x^N. \tag{2.4}$$

Obcięcie szeregu do N pierwszych wyrazów wprowadza błąd.

Błędy zaokrągleń (*round-off errors*)

Występują podczas wykonywania operacji arytmetycznych na liczbach zmiennoprzecinkowych, gdy wynik operacji nie może być dokładnie reprezentowany w dostępnej precyzji i musi zostać zaokrąglony.

Przykład: Dzielenie 1/3 daje nieskończone rozwinięcie dziesiętne (0.333…). Jeśli komputer przechowuje tylko pewną liczbę cyfr, wynik zostanie zaokrąglony, np. do 0.333. Błędy te kumulują się w długich ciągach obliczeń.

2.2. Miary błędów

Niezwykle istotnym problemem w obliczeniach numerycznych jest występowanie błędów i ich estymacja. Zasadniczo błędy możemy wyrażać w postaci bezwzględnej (*absolut error*) i względnej (*relative error*).

Błąd bezwzględny (Δ): Różnica między wartością dokładną A a wartością przybliżoną a:

$$\Delta = |A - a| \tag{2.5}$$

Błąd względny (δ): Stosunek błędu bezwzględnego do wartości dokładnej

$$\delta = \frac{|A-a|}{|A|} \text{ dla } A \neq 0 \tag{2.6}$$

często wyrażany w procentach:

$$\%\,\delta = \frac{|A-a|}{|A|}\cdot 100\,\%\ \text{ dla }\ A \neq 0. \tag{2.7}$$

Błąd względny jest zazwyczaj bardziej informatywny, ponieważ wskazuje na "jakość" przybliżenia w stosunku do wielkości samej liczby.

W wielu przypadkach jednak wartość dokładna nie jest znana, stąd konieczność estymacji błędów na podstawie innych kryteriów, o których będzie mowa w późniejszej części.

Propagacja błędów: Błędy nie pozostają statyczne; często propagują się (przenoszą i kumulują) w kolejnych etapach obliczeń. Małe błędy początkowe mogą prowadzić do znacznie większych błędów w końcowym wyniku, szczególnie w przypadku niestabilnych numerycznie algorytmów.

Kumulacja błędów: Jeśli wiele operacji jest wykonywanych jedna po drugiej, błędy z poprzednich operacji sumują się i mogą wpływać na ostateczny wynik.

Niestabilność numeryczna: Algorytm jest niestabilny numerycznie, jeśli małe błędy danych wejściowych lub błędy powstałe w trakcie obliczeń szybko rosną, prowadząc do dużych zniekształceń wyników.

Analiza błędów: Celem analizy błędów jest oszacowanie wielkości błędów w wynikach obliczeń i zrozumienie, jak wpływają one na wiarygodność uzyskanych rezultatów.

Ocena dokładności: Określenie, ile cyfr znaczących w wyniku jest wiarygodnych.

Minimalizacja błędów: Poszukiwanie algorytmów i metod obliczeniowych, które są mniej wrażliwe na propagację błędów lub które generują mniejsze błędy obcięcia i zaokrągleń. Może to obejmować:

- Przekształcanie wzorów, aby unikać operacji, które są szczególnie podatne na błędy (np. odejmowanie bardzo bliskich sobie liczb, co prowadzi do utraty precyzji).
- Zwiększanie precyzji obliczeń (np. użycie liczb zmiennoprzecinkowych podwójnej precyzji).
- Wybór odpowiedniej metody numerycznej, która jest numerycznie stabilna dla danego problemu.

Analiza wrażliwości: Badanie, jak zmiany w danych wejściowych (lub małe błędy) wpływają na wyniki końcowe.

W praktyce, inżynierowie i naukowcy muszą być świadomi tych błędów, aby prawidłowo interpretować wyniki uzyskane z symulacji i obliczeń numerycznych.

3. Rozwiązywanie układów równań liniowych

Metody rozwiązywania równań liniowych można ogólnie podzielić na dwie kategorie: techniki bezpośrednie i iteracyjne. Techniki bezpośrednie dostarczają rozwiązania w skończonej, przewidywalnej liczbie kroków. Dodatkowo, w przypadku tych metod rozwiązanie jest dokładne obarczone jedynie błędem zaokrągleń i reprezentacji liczb. Natomiast techniki iteracyjne dają rozwiązanie w nieprzewidywalnej liczbie kroków. Im większa liczba iteracji, tym dokładniejsze staje się rozwiązanie. Przykładem pierwszej techniki rozwiązywania liniowego układu równań algebraicznych jest *metoda eliminacji Gaussa*. Drugą grupę reprezentuje iteracyjna *metoda Gaussa-Siedla*.

Rozdział ten dotyczyć będzie rozwiązywania układu równań w postaci:

$$\begin{cases} a_{11}\cdot x_1+a_{12}\cdot x_2+\ldots+a_{1n}\cdot x_n=b_1 \\ a_{21}\cdot x_1+a_{22}\cdot x_2+\ldots+a_{2n}\cdot x_n=b_2 \\ \vdots \\ a_{n1}\cdot x_1+a_{n2}\cdot x_2+\ldots+a_{nn}\cdot x_n=b_n \end{cases}. \tag{3.1}$$

Ten sam układ równań możemy zapisać w postaci macierzowej:

$$\boldsymbol{A}\cdot\boldsymbol{x}=\boldsymbol{b} \tag{3.2}$$

gdzie $\boldsymbol{A}$ jest rzeczywistą, nieosobliwą macierzą ($\det(\boldsymbol{A})=|\boldsymbol{A}|\neq 0$) o wymiarze $n\times n$, $\boldsymbol{b}$ jest rzeczywistym wektorem o długości n oraz $\boldsymbol{x}$ jest wektorem niewiadomych o długości n. W niniejszych materiałach operator $|*|$ będzie oznaczał wyznacznik jeśli dotyczy macierzy lub wartość bezwzględną jeśli będzie dotyczył wartości skalarnej. W równaniach, wielkości wektorowe oznaczane będą za pomocą symboli zapisanych pogrubioną czcionką lub wykorzystując zapis wskaźnikowy.

Macierz $\boldsymbol{A}$, wektory $\boldsymbol{x}$ i $\boldsymbol{b}$ można przedstawić w formie jawnej:

$$\boldsymbol{A}=\begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \vdots & \vdots & \ldots & \vdots \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{bmatrix},\ \boldsymbol{x}=\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix},\ \boldsymbol{b}=\begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}. \tag{3.3}$$

3.1. Uwarunkowanie zadania

Zasadniczym czynnikiem decydującym o wyborze metody rozwiązania jest uwarunkowanie układu równań. Stopień uwarunkowania układu sprowadza się do obliczenie współczynnika uwarunkowania macierzy $\boldsymbol{A}$, który nazywany jest $\mathrm{cond}(\boldsymbol{A})$. Macierze, których wartość współczynnika uwarunkowania jest mała nazywamy macierzami dobrze uwarunkowanymi (*well conditioned matrix*), natomiast macierze, których wartość współczynnika uwarunkowania jest duża nazywamy macierzami źle uwarunkowanymi (*ill conditioned matix*).

Współczynnik uwarunkowania macierzy obliczany jest jako:

$$\mathrm{cond}(\boldsymbol{A}) = \|\boldsymbol{A}\| \cdot \|\boldsymbol{A}^{-1}\| \tag{3.4}$$

gdzie $\|\boldsymbol{A}\|$ jest normą macierzy $\boldsymbol{A}$ natomiast $\|\boldsymbol{A}^{-1}\|$ jest normą macierzy odwrotnej do $\boldsymbol{A}$. Do obliczenia współczynnika uwarunkowania możemy zastosować trzy rodzaje norm:

Norma kolumnowa (tzw. norma pierwsza)

$$\|\boldsymbol{A}\|_1 = \max_{1 \leqslant i \leqslant n} \sum_{i=1}^{n} |a_{ij}| \tag{3.5}$$

gdzie n oznacza wymiar macierzy oraz i, j oznaczają odpowiednio numery wierszy i kolumn.

Norma spektralna (tzw. norma druga)

$$\|\boldsymbol{A}\|_2 = \max_{1 \leqslant i \leqslant n} |\lambda_i| \tag{3.5}$$

gdzie λ_i oznacza i-tą wartość własną macierzy $\boldsymbol{A}$. Wartości własne λ_i skojarzone z wektorami własnymi $\boldsymbol{v_i}$ spełniają równania:

$$\boldsymbol{A} \cdot \boldsymbol{v_i} = \lambda_i \cdot \boldsymbol{v_i} \tag{3.6}$$

Co jest równoznaczne równaniu

$$(\boldsymbol{A} - \lambda_i \boldsymbol{I}) \cdot \boldsymbol{v_i} = \boldsymbol{0} \tag{3.7}$$

gdzie $\boldsymbol{I}$ jest macierzą jednostkową. Układ równań (3.7) ma rozwiązanie zerowe jeśli:

$$|\boldsymbol{A} - \lambda \boldsymbol{I}| = 0 \tag{3.8}$$

Z równania (3.8) można policzyć wartości własne, wyznaczając pierwiastki funkcji wielomianowej stopnia n. Znając wartości własne, wektory własne możemy

wyznaczyć rozwiązują układ równań (3.7) dla każdej wyznaczonej wcześniej wartości własnej (nie jest to konieczne dla obliczenia normy spektralnej).

Norma spektralna może być również wyrażana za pomocą promienia spektralnego ρ iloczynu macierzy współczynników układu równań $\boldsymbol{A}^T \cdot \boldsymbol{A}$:

$$\|\boldsymbol{A}\|_2 = \sqrt{\rho(\boldsymbol{A}^T \cdot \boldsymbol{A})} \tag{3.9}$$

Promień spektralny macierzy $\boldsymbol{M} = \boldsymbol{A}^T \cdot \boldsymbol{A}$ definiowany jest jako maksymalna wartość własna macierzy $\boldsymbol{M}$:

$$\rho(\boldsymbol{M}) = \max_{1 \leqslant i \leqslant n} |\lambda_i|. \tag{3.10}$$

Norma wierszowa (tzw. norma nieskończoność)

$$\|\boldsymbol{A}\|_\infty = \max_{1 \leqslant j \leqslant n} \sum_{i=1}^{n} |a_{ij}|. \tag{3.11}$$

Zadanie 3.1. Wyznacz stopień uwarunkowania macierzy według normy wierszowej, kolumnowej i spektralnej (ręcznie):

$$\boldsymbol{A} = \begin{bmatrix} 1 & 2 \\ 2 & 3 \end{bmatrix}.$$

Zadanie 3.2. Wyznacz stopnie uwarunkowania macierzy Hilberta, której elementy są zdefiniowane następująco: $h_{ij} = \frac{1}{i + j - 1}$, dla rozmiaru $n = 8$, $(i, j = 1,2,\ldots,n)$, (wykorzystaj odpowiednie biblioteki języka Python).

3.2. Metody dokładne

Metody dokładne pozwalają na znalezienie rozwiązania dokładnego w określonej liczbie kroków. Rozwiązanie jest dokładne pod warunkiem że obliczenia nie są obarczone błędami zaokrągleń. W praktyce obliczenia wykonywane są przy pomocy komputerów co wprowadza do rozwiązania błąd wynikający z reprezentacji liczb rzeczywistych w systemie komputerowym i w konsekwencji zaokrąglenia wyniku. Metody dokładne stosowane są w praktyce do rozwiązywanie niezbyt dużych układów równań, ze względu na wymaganą dużą ilość działań arytmetycznych.

Metoda Cramera

Metoda Cramera jest najbardziej znaną metodą rozwiązywania układu równań w algebrze liniowej. Algorytm metody Cramera wymaga obliczenia wyznacznika macierzy współczynników układu równań (3.2) oraz wyznaczników macierzy przekształconej poprzez zastąpienie kolumny w macierzy wyznaczników wektorem prawej strony co wyraża wzór:

$$x_i = \frac{|\boldsymbol{A}|}{|\boldsymbol{A}_i|} \text{ dla } i = 1,2,\ldots n \tag{3.12}$$

gdzie $\boldsymbol{A}_i$ oznacza przekształconą macierz $\boldsymbol{A}$ w której kolumnę i-tą zastąpiono wektorem prawej strony $\boldsymbol{b}$.

Zatem aby rozwiązać układ równań należy policzy $n+1$ wyznaczników. Obliczenie wyznacznika macierzy o rozmiarze $n \times n$ wymaga $(n-1)n!$ operacji, co w sumie prowadzi do $(n+1)(n-1)(n)! \approx n^{n+2}$ działań. Z tego powodu metoda jest w praktyce rzadko wykorzystywana ze względu na wysoką złożoność czasową.

Metoda eliminacji Gaussa

Metoda eliminacji Gaussa polega na iteracyjnym, metodycznym przekształceniu układu równań $\boldsymbol{A} \cdot \boldsymbol{x} = \boldsymbol{b}$ do postaci, w której macierz współczynników układu równań jest macierzą górno-trójkątną $\boldsymbol{A}^{[n]} \cdot \boldsymbol{x} = \boldsymbol{b}^{[n]}$. Kolejne formy układu równań będą reprezentowane przez indeks $[k]$ wskazujący na postać układu równań w kolejnych etapach jego przekształcania. Tak przekształcony układ równań daje się prosto rozwiązać metodą podstawień wstecznych.

Rozwiązanie układu równań pozostanie niezmienione, jeśli zostanie wykonana dowolna z następujących operacji:

• Mnożenie lub dzielenie dowolnego równania przez stałą,

• Zastąpienie dowolnego równania sumą lub różnicą tego równania i dowolnego innego równania.

Eliminacja Gaussa to po prostu sekwencja tych podstawowych operacji wierszowych w celu przekształcenia macierzy współczynników $\boldsymbol{A}$ do postaci górno-trójkątnej.

Korzystając z powyższego schematu, wyjściowa postać rozwiązywanego układu równań zostanie oznaczona indeksem $k = 1$, i przyjmie postać:

$$\boldsymbol{A}^{[1]} \cdot \boldsymbol{x} = \boldsymbol{b}^{[1]} \tag{3.13}$$

gdzie

$$\begin{cases} a_{11}^{[1]} \cdot x_1 + a_{12}^{[1]} \cdot x_2 + \ldots + a_{1n}^{[1]} \cdot x_n = b_1^{[1]} \\ a_{21}^{[1]} \cdot x_1 + a_{22}^{[1]} \cdot x_2 + \ldots + a_{2n}^{[1]} \cdot x_n = b_2^{[1]} \\ \vdots \\ a_{n1}^{[1]} \cdot x_1 + a_{n2}^{[1]} \cdot x_2 + \ldots + a_{nn}^{[1]} \cdot x_n = b_n^{[1]} \end{cases} . \tag{3.14}$$

W pierwszym kroku eliminujemy zmienną x_1 z wierszy o 2 do n, co oznacza wyzerowanie współczynników w pierwszej kolumnie macierzy $\boldsymbol{A}^{[1]}$ pod elementem $a_{11}^{[1]}$

(tzw. elementem wiodącym w kroku 1). Zakładając, że $a_{11}^{[1]} \neq 0$, wykonujemy dla każdego wiersza $i = 2,3 \ldots n$ operacje:

- obliczamy współczynnik $d_{i1}^{[1]} = \frac{a_{i1}^{[1]}}{a_{11}^{[1]}}$
- następnie pierwszy wiersz układu mnożymy przez współczynnik $d_{i1}^{[1]}$ i odejmujemy od i-tego wiersza układu.

W ten sposób otrzymujemy zmodyfikowany układu równań w kolejnej iteracji $k = 2$, wyrażony poprzez przekształconą macierz współczynników $\boldsymbol{A}^{[2]}$ i przekształcony wektor prawej strony $\boldsymbol{b}^{[2]}$:

$$\boldsymbol{A}^{[2]} = \begin{bmatrix} a_{11}^{[2]} & a_{12}^{[2]} & \ldots & a_{1n}^{[2]} \\ 0 & a_{22}^{[2]} & \ldots & a_{2n}^{[2]} \\ \vdots & \vdots & \ldots & \vdots \\ 0 & a_{n2}^{[2]} & \ldots & a_{nn}^{[2]} \end{bmatrix}, \quad \boldsymbol{b} = \begin{bmatrix} b_1^{[2]} \\ b_2^{[2]} \\ \vdots \\ b_n^{[2]} \end{bmatrix}. \tag{3.15}$$

W drugim etapie eliminujemy zmienną x_2 z wierszy o 3 do n, co oznacza wyzerowanie współczynników w drugiej kolumnie macierzy $\boldsymbol{A}^{[2]}$, pod elementem $a_{22}^{[2]}$, a więc element $a_{22}^{[2]}$ staje się elementem wiodącym.

Zakładając, że $a_{22}^{[2]} \neq 0$, wykonujemy te same operacje co w pierwszym kroku, tym razem dla wierszy $i = 3,4 \ldots n$:

- obliczamy współczynnik $d_{i2}^{[2]} = \frac{a_{i2}^{[2]}}{a_{22}^{[2]}}$
- następnie drugi (k-ty) wiersz układu mnożymy przez współczynnik $d_{i2}^{[2]}$ i odejmujemy od i-tego wiersza macierzy układu.

Po drugim etapie otrzymujemy następujące postacie macierzy $\boldsymbol{A}^{[3]}$ i wektora $\boldsymbol{b}^{[3]}$:

$$\boldsymbol{A}^{[3]} = \begin{bmatrix} a_{11}^{[3]} & a_{12}^{[3]} & \ldots & a_{1n}^{[3]} \\ 0 & a_{22}^{[3]} & \ldots & a_{2n}^{[3]} \\ \vdots & \vdots & \ldots & \vdots \\ 0 & 0 & \ldots & a_{nn}^{[3]} \end{bmatrix}, \quad \boldsymbol{b} = \begin{bmatrix} b_1^{[3]} \\ b_2^{[3]} \\ \vdots \\ b_n^{[3]} \end{bmatrix}. \tag{3.16}$$

W kolejnych etapach aż do $k = n$ wykonujemy dokładnie te same operacje, w konsekwencji otrzymując układ równań z macierzą współczynników w postaci górno-trójkątnej:

$$\boldsymbol{A}^{[n]}=\begin{bmatrix} a_{11}^{[n]} & a_{12}^{[n]} & \dots & a_{1n}^{[n]} \\ 0 & a_{22}^{[n]} & \dots & a_{2n}^{[n]} \\ \vdots & \vdots & \dots & \vdots \\ 0 & 0 & \dots & a_{nn}^{[n]} \end{bmatrix},\ \boldsymbol{b}=\begin{bmatrix} b_1^{[n]} \\ b_2^{[n]} \\ \vdots \\ b_n^{[n]} \end{bmatrix}. \tag{3.17}$$

Metoda podstawień wstecznych, polega na wyznaczeniu z ostatniego równania niewiadomej x_n, podstawienia tej wartości do równania $n-1$ i wyliczenia z tego równania zmiennej x_{n-1}, następnie podstawiamy wyznaczone wartości zmiennych x_n i x_{n-1} do równania $n-3$ i wyznaczmy z tego równania wartość niewiadomej x_3 itd. aż do równania 1. Algorytm podstawień wstecznych możemy przedstawić w następujących krokach:

- podstaw za $x_n=\dfrac{b_n^{[n]}}{a_{nn}^{[n]}}$
- ustal indeks $i=n-1$
- oblicz niewiadomą x_i ze wzoru $x_i=\dfrac{b_i^{[i]}-\sum_{j=i+1}^{n} a_{ij}^{[i]}x_j}{a_{ii}^{[i]}}$
- jeżeli $i>1$, to zmniejsz i o 1, i przejdź do punktu poprzedniego, w przeciwnym razie zakończ obliczenia.

Metoda eliminacji Gaussa z wyborem elementu wiodącego

W metodzie eliminacji Gaussa zakłada się, że element wiodący w danym kroku jest różny od zera, w pewnych jednak sytuacjach, mimo że macierz współczynników układu równań nie jest osobliwa, napotkać można na zerową wartość na diagonalnej co prowadzi do dzielenia przez zero. Rozwiązaniem tego problemu jest przekształcenie układu równań przez zamianę wierszy i/lub kolumn w taki sposób aby element wiodący był niezerowy. Istnieją dwa podejścia do rozwiązania tego problemu: *metoda z częściowym wyborem elementu wiodącego* lub *metoda z pełnym wyborem elementu wiodącego*.

Metoda z częściowym wyborem elementu wiodącego polega na tym, że w każdym etapie przekształcania układu równań szukany jest element wiodący wierszami lub kolumnami. W przypadku poszukiwania z zamianą wierszy w k-tym etapie wykonywane są następujące kroki:

- szukany jest współczynnik o największym module w k-tej kolumnie

$$|a_{mk}^{[k]}| = \max_{j=k,k+1\ldots n} |a_{jk}^{[k]}| \tag{3.18}$$

- wiersze m i k zamieniane są miejscami.

Analogicznie postępuje się w przypadku częściowego wyboru elementu wiodącego z zamianą kolumn, wówczas poszukiwany jest współczynnik o największym module w k-tej kolumnie i przestawiane są ze sobą odpowiednie kolumny.

Metoda z pełnym wyborem elementu wiodącego polega na tym że w k-tym etapie eliminacji poszukiwany jest współczynnik o największym module spośród wszystkich elementów a_{ij} pod-macierzy $\boldsymbol{A}^{[k]}$:

$$|a_{mk}^{[k]}| = \max_{k \leqslant i,\, j \leqslant n} |a_{ij}^{[k]}| \tag{3.18}$$

Następnie zamieniane są zarówno kolumny jak i wiersze.

3.3. Metody iteracyjne

Metody iteracyjne mają mniejszą złożoność obliczeniową niż metody dokładne, są bardziej efektywne w rozwiązywaniu dużych układów równań, dając rozwiązanie przybliżone. W niniejszym rozdziale będą omawiane metody iteracji prostej, polegające na przekształceniu układu równań $\boldsymbol{A}\cdot\boldsymbol{x}=\boldsymbol{b}$ w układ równoważny postaci

$$\boldsymbol{x}=\boldsymbol{M}\cdot\boldsymbol{x}+\boldsymbol{c} \tag{3.19}$$

gdzie $\boldsymbol{M}$, $\boldsymbol{c}$ są rzeczywistą macierzą i rzeczywistym wektorem o rozmiarze $n\times n$ i n, odpowiednio.

Układ równań (3.19) jest następnie iteracyjnie rozwiązywany. Metody iteracyjne możemy stosować do układów równań spełniających warunki:

- macierz $\boldsymbol{A}$ jest macierzą nieosobliwą $|\boldsymbol{A}|\neq 0$, z niezerowymi elementami na głównej przekątnej $a_{ii}\neq 0$.

- warunkiem koniecznym i wystarczającym zbieżności metod iteracji prostej jest aby promień spektralny macierzy $\boldsymbol{M}$ spełniał warunek $\rho(\boldsymbol{M})<1$.

W metodach iteracji prostej tworzymy ciąg przybliżeń $x^{(k)}$ rozwiązania dokładnego $\bar{\boldsymbol{x}}$

$$\boldsymbol{x}^{(k+1)}=\boldsymbol{M}\cdot\boldsymbol{x}^{(k)}+\boldsymbol{c} \quad k=1,2,\ldots \tag{3.20}$$

w taki sposób, że:

$$\lim_{k\to\infty} \boldsymbol{x}^{(k)}=\bar{\boldsymbol{x}} \tag{3.21}$$

Postać macierzy $\boldsymbol{M}$ i wektora $\boldsymbol{c}$ zależy od wybranej metody.

Metoda Jacobiego

Jedną z metod iteracji prostej jest metoda Jacobiego, polegająca na rozkładzie macierzy $\boldsymbol{A}$ na sumę trzech macierzy w taki sposób, że:

$$\boldsymbol{A}=\boldsymbol{L}+\boldsymbol{D}+\boldsymbol{U} \tag{3.22}$$

gdzie: $\boldsymbol{L}$ jest macierzą dolno-trójkątną z zerami na diagonali (pod-diagonalną), $\boldsymbol{D}$ jest macierzą diagonalną (a_{ii}<>0 dla każdego $i=1,2\ldots n$), $\boldsymbol{U}$ jest macierzą górno-trójkątną z zerami na diagonali (nad-diagonalną).

Układ równań $\boldsymbol{A}\cdot\boldsymbol{x}=\boldsymbol{b}$ z wprowadzonym rozkładem (3.22) przyjmuje postać:

$$(\boldsymbol{L}+\boldsymbol{D}+\boldsymbol{U})\cdot\boldsymbol{x}=\boldsymbol{b}. \tag{3.23}$$

Po przekształceniu otrzymujemy

$$\boldsymbol{D}\cdot\boldsymbol{x}=-(\boldsymbol{L}+\boldsymbol{U})\cdot\boldsymbol{x}+\boldsymbol{b} \tag{3.23}$$

stąd

$$\boldsymbol{x}=-\boldsymbol{D}^{-1}(\boldsymbol{L}+\boldsymbol{U})\cdot\boldsymbol{x}+\boldsymbol{D}^{-1}\cdot\boldsymbol{b}. \tag{3.24}$$

Macierz $\boldsymbol{D}$ jako macierz diagonalna jest łatwo odwracalna, aby policzyć macierz odwrotną $\boldsymbol{D}^{-1}$ wystarczy odwrócić elementy na diagonali.

Z równania (3.24) ostatecznie otrzymujemy wzór iteracyjny metody Jacobiego

$$\boldsymbol{x}^{(k+1)}=-\boldsymbol{D}^{-1}(\boldsymbol{L}+\boldsymbol{U})\cdot\boldsymbol{x}^{(k)}+\boldsymbol{D}^{-1}\cdot\boldsymbol{b} \tag{3.24}$$

gdzie macierz

$$\boldsymbol{M}_J=-\boldsymbol{D}^{-1}(\boldsymbol{L}+\boldsymbol{U}) \tag{3.25}$$

nazywamy macierzą Jacobiego.

Aby wyznaczyć błąd rozwiązania i warunek zakończenia iteracji nie możemy posłużyć się definicją błędu (2.5-2.7) ze względu, że rozwiązanie dokładne nie jest znane. W takim przypadku możemy wykorzystać miarę błędu będącą maksymalną bezwzględną różnicą między rozwiązaniami w kolejnych iteracjach:

$$\Delta=\max\left|\boldsymbol{x}^{(k+1)}-\boldsymbol{x}^{(k)}\right| \tag{3.26}$$

lub miarą względną (jeśli rozwiązania jest niezerowe):

$$\delta = \max \frac{|\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)}|}{|\boldsymbol{x}^{(k+1)}|}. \tag{3.27}$$

Miary błędów w takim przypadku opierają się na założeniu zbieżności metody do dokładnego rozwiązania a więc w kolejnych iteracjach różnice rozwiązań będą dążyć do zera.

Algorytm metody Jacobiego możemy przedstawić zatem w postaci kroków:

1. Wprowadź dane wejściowe: $\boldsymbol{A}$, $\boldsymbol{b}$, ε - dokładność obliczeń.
2. Rozłóż macierz $\boldsymbol{A}$ na trzy macierze $\boldsymbol{L}$, $\boldsymbol{D}$, $\boldsymbol{U}$.
3. Oblicz promień spektralny (3.10) macierzy Jacobiego (3.25), jeśli jest mniejszy od 1 przejdź do następnego kroku, w przeciwnym przypadku zakończ procedurę.
4. Przyjmij wektor startowy $\boldsymbol{x}^{(0)}$ i $k = 0$.
5. Wyznacz wektor $\boldsymbol{x}^{(k+1)}$ korzystając ze wzoru (3.24).
6. Jeśli warunek $\max |\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)}| < \varepsilon$ jest spełniony uznaj $\boldsymbol{x}^{(k+1)}$ za rozwiązanie i zakończ procedurę w przeciwnym wypadku zwiększ k o 1 i przejdź do punktu 5.

Metoda Gaussa-Siedla

W metodzie Gaussa-Siedla równanie (3.23) przekształcane jest do postaci:

$$(\boldsymbol{L} + \boldsymbol{D}) \cdot \boldsymbol{x} = -\boldsymbol{U} \cdot \boldsymbol{x} + \boldsymbol{b} \tag{3.28}$$

stąd

$$\boldsymbol{x} = -(\boldsymbol{L} + \boldsymbol{D})^{-1} \cdot \boldsymbol{U} \cdot \boldsymbol{x} + (\boldsymbol{L} + \boldsymbol{D})^{-1} \cdot \boldsymbol{b}. \tag{3.29}$$

W tym przypadku również macierz $(\boldsymbol{L} + \boldsymbol{D})$ jest macierzą łatwo odwracalną i w wyniku otrzymujemy również macierz dolno-trójkątną.

Wzór iteracyjny w metodzie Gaussa-Siedla przyjmuje postać:

$$\boldsymbol{x}^{(k+1)} = -(\boldsymbol{L} + \boldsymbol{D})^{-1} \cdot \boldsymbol{U} \cdot \boldsymbol{x}^{k} + (\boldsymbol{L} + \boldsymbol{D})^{-1} \cdot \boldsymbol{b}. \tag{3.30}$$

Należy zauważyć, że w metodzie Gaussa-Siedla posługujemy się najbardziej aktualnym rozwiązaniem, co możemy zapisać w formia jawnej jako:

$$x_i^{(k+1)} = -\frac{1}{a_{ii}} \cdot \left(\sum_{j<i} a_{ij}^{k+1} x_j + \sum_{j>i} a_{ij}^{k} x_j \right) + \frac{b_i}{a_{ii}} \text{ dla } i = 1,2 \ldots n \tag{3.31}$$

co oznacza, że w każdej kolejnej iteracji mamy do dyspozycji rozwiązanie $\boldsymbol{x}^{k}$ i częściowo $\boldsymbol{x}^{(k+1)}$.

Macierz

$$\boldsymbol{M}_{GS} = -(\boldsymbol{L} + \boldsymbol{D})^{-1} \cdot \boldsymbol{U} \tag{3.32}$$

jest nazywana macierzą Gaussa-Siedla. Procedurę Gaussa-Siedla możemy przedstawić w postaci:

1) wprowadź dane wejściowe: $\boldsymbol{A}$, $\boldsymbol{b}$, ε - dokładność obliczeń,
2) rozłóż macierz $\boldsymbol{A}$ na trzy macierze $\boldsymbol{L}$, $\boldsymbol{D}$, $\boldsymbol{U}$,
3) oblicz promień spektralny (3.10) macierzy Gaussa-Siedla (3.30), jeśli jest mniejszy od 1 przejdź do następnego kroku, w przeciwnym przypadku zakończ procedurę,
4) przyjmij wektor startowy $\boldsymbol{x}^{(0)}$ i $k = 0$,
5) wyznacz wektor $\boldsymbol{x}^{(k+1)}$ korzystając ze wzoru (3.31),
6) jeśli warunek $\max |\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)}| < \varepsilon$ jest spełniony uznaj $\boldsymbol{x}^{(k+1)}$ za rozwiązanie i zakończ procedurę w przeciwnym wypadku zwiększ k o 1 i przejdź do punktu 5.

Uwaga

W praktyce obliczenie promienia spektralnego macierzy, zwłaszcza o dużych rozmiarach, jest zadaniem wymagającym obliczeniowo. Alternatywnie możemy posłużyć się twierdzeniem 3.1.

Twierdzenie 3.1: *Jeżeli macierz $\boldsymbol{A}$ spełnia jeden z warunków:*

1) $|a_{ii}| > \sum_{i \neq j} a_{ij}$ dla $i = 1,2 \ldots n$ *(mocne kryterium wierszy)*
2) $|a_{ii}| > \sum_{i \neq j} a_{ij}$ dla $j = 1,2 \ldots n$ *(mocne kryterium kolumn)*

to metody Jacobiego i Gaussa-Siedla są zbieżne.

Kryteria te oznaczają macierz tzw. skupioną gdzie największe bezwzględne wartości skupione są na diagonali. W celu doprowadzenia do takiej postaci macierzy można skorzystać z metod eliminacji Gaussa z wyborem elementu wiodącego.

Ostatecznie można ostrożnie prowadzić proces iteracji sprawdzając jak zmienia się błąd rozwiązania, w przypadku jego zwiększania należy procedurę przerwać.

Zadanie 3.3. Mając dany układ równań

$$\begin{cases} 1\cdot x_1+2\cdot x_2+3\cdot x_3=1 \\ 2\cdot x_1+1\cdot x_2+3\cdot x_3=2 \\ 3\dot{x}_1+2\cdot x_2+1\cdot x_n=0 \end{cases}$$

rozwiąż go (ręcznie) za pomocą metod dokładnych:

- metody Cramera,
- metody eliminacji Gaussa

i metod iteracyjnych:

- metody Jacobiego,
- metody Gausa-Siedla.

Zadanie 3.4. Zaimplementuj metodę Jacobiego do rozwiązywania układu n-równań z wykorzystaniem odpowiednich bibliotek języka Python.

4. Rozwiązywanie równań nieliniowych

Dla większości spotykanych w praktyce równań nieliniowych nie istnieją metody dokładne ich rozwiązania, w takim przypadku pozostają jedynie metody przybliżone. Metody przybliżone są zawsze metodami iteracyjnymi co oznacza, że konstruujemy ciąg kolejnych przybliżeń x_n zbieżnych do rozwiązania dokładnego $\bar{x}$.

Rozdział ten dotyczy rozwiązywania równań nieliniowych zdefiniowanych funkcją $f(x)$:

$$f(x)=0 \tag{4.1}$$

metodami przybliżonymi opartymi na twierdzeniu Bolzano-Cauchy'ego:

Twierdzenie 4.1: *Jeżeli funkcja $f(x)$ określona i ciągła w przedziale domkniętym $[a,b]$ przyjmuje w punktach a i b różne znaki to wewnątrz przedziału $[a,b]$ posiada co najmniej jeden pierwiastek równania $f(x)=0$.*

Aby ciąg kolejnych przybliżeń x_n był zbieżny do rozwiązania dokładnego $\bar{x}$ musi spełniać nierówność dla dostatecznie dużych n:

$$|\bar{x}-x_{n+1}|\leqslant C|\bar{x}-x_n|^p \tag{4.2}$$

Gdzie C jest stałą zależną od $f(x)$, p jest wykładnikiem zbieżności. Im mniejsza jest wartość C i im większy jest wykładnik p tym ciąg kolejnych przybliżeń x_n jest szybciej zbieżny do rozwiązania dokładnego $\bar{x}$. Wartość p zależy od metody rozwiązania i mówimy o metodzie rzędu p. Jeżeli $p=1$ to metoda jest zbieżna liniowo, natomiast jeśli $p>1$ to metoda jest zbieżna ponad-liniowo, w szczególności gdy $p=2$ mówimy o kwadratowej zbieżności metody.

Błąd metod iteracyjnych definiujący warunek zakończenia iteracji można określać za pomocą kryterium wartości funkcji:

$$f(x_n)\leqslant\varepsilon \tag{4.3}$$

gdzie ε jest żądaną dokładnością rozwiązania, lub za pomocą różnicy między kolejnymi przybliżeniami w formie bezwzględnej:

$$|x_{n+1}-x_n|\leqslant\varepsilon \tag{4.4}$$

lub względnej:

$$\frac{|x_{n+1}-x_n|}{|x_{n+1}|}\leqslant\varepsilon \text{ dla } x_{n+1}\neq 0. \tag{4.5}$$

Najczęściej stosowana jest kombinacja tych warunków.

4.1. Metoda bisekcji

Metoda bisekcji polega na zawężaniu przedziału izolacji pierwiastka poprzez połowienie przedziału. Jeżeli środek przedziału nie jest szukanym miejscem zerowym, to połowiony jest przedział na krańcach którego wartości funkcji f są przeciwnych znaków.

Gwarancją zbieżności metody bisekcji jest aby funkcja $f(x)$ w przedziale $[a,b]$ spełniała warunki:

- była ciągła,
- była różnych znaków na krańcach przedziału.

Zatem algorytm metody bisekcji można zapisać w następujący sposób:

1) przyjmij $k=0$, $x_0=a$, $y_0=b$ i ε - błąd rozwiązania,

2) wyznacz $z_k=\frac{x_k+y_k}{2}$,

3) jeśli $|f(z_k)|\leqslant\varepsilon$ lub $|y_k-x_k|\leqslant\varepsilon$ zakończ procedurę, w przeciwnym wypadku przejdź do następnego punktu,

4) jeśli $f(x_k)f(z_k)>0$ to wyznacz $x_{k+1}=z_k$, $y_{k+1}=y_k$, w przeciwnym wypadku $x_{k+1}=x_k$, $y_{k+1}=z_k$,

5) powiększ k o 1 i przejdź do punktu 2.

Po zakończeniu procedury wyznaczonym miejscem zerowym jest wartość z_k.

4.2. Metoda iteracji prostej

Metoda iteracji prostej, znana również jako metoda iteracji punktu stałego, to technika numeryczna służąca do znajdowania rozwiązań równań poprzez iteracyjne przybliżanie rozwiązania.

Metoda iteracji prostej opiera się na koncepcji punktu stałego. Punkt stały funkcji $g(x)$ to taki punkt x, dla którego $g(x)=x$. Aby znaleźć pierwiastek równania $f(x)=0$, przekształcamy je do postaci $x=g(x)$ i wykonujemy iteracje, aż x_n będzie dostatecznie blisko punktu stałego.

Przekształcenie równania $f(x)=0$ może następować przez jego re-aranżację lub dodanie do obu stron równania funkcji x. Dobór funkcji $g(x)$ ma kluczowe znaczenie dla zbieżności metody, gdyż nie wszystkie przekształcenia gwarantują zbieżność metody.

Wyznaczenie miejsca zerowego $\bar{x}$ funkcji $f(x)$ w przedziale $[a,b]$ metodą iteracji prostej, sprowadza się do przekształcenia funkcji $f(x)$ do postaci:

$$g(x)=x \text{ gdzie } g(x)=x+\alpha f(x). \tag{4.6}$$

Metoda polega na tworzeniu ciągu iteracyjnego x_n dla $n=1,2\ldots$ zgodnie z regułą

$$x_{n+1}=x_n+\alpha f(x_n). \tag{4.7}$$

Zbieżność procesu iteracyjnego do granicy $\bar{x}$ w $[a,b]$ zależy od właściwego wyboru parametru α oraz spełnienia warunków:

$$\bar{x}\in[a,b],\ f(a)\cdot f(b)<0,\ |g'(x)|<1 \text{ dla } x\in[a,b]. \tag{4.8}$$

4.3. Metoda Newtona-Raphsona

Metoda Newtona-Raphsona, inaczej nazywana metodą stycznych opiera się na wyznaczaniu stycznych do wykresu badanej funkcji $f(x)$ w punktach kolejnych przybliżeń. Zbieżność metody Newtona-Raphsona gwarantuje spełnienie w przedziale izolacji $[a,b]$ następujących warunków:

1) funkcja jest ciągła w przedziale $[a,b]$,
2) pierwsza i druga pochodna funkcji istnieją i są ciągłe w przedziale $[a,b]$,
3) iloczyn wartości funkcji na krańcach przedziału spełnia warunek $f(a)\cdot f(b)<0$, co oznacza zmianę znaku funkcji $f(x)$ w przedziale $[a,b]$,
4) pierwsza i druga pochodna funkcji $f(x)$ mają stały znak w przedziale izolacji $[a,b]$, co oznacza, że w tym przedziale funkcja $f(x)$ nie ma ekstremów lokalnych i punktów przegięcia,
5) punktem startowym obliczeń powinien być ten koniec przedziału $[a,b]$ w którym funkcja $f(x)$ przyjmuje ten sam znak co jej druga pochodna, czyli spełniony jest warunek $f(x_0)\cdot f''(x_0)>0$ dla $x_0=a$ lub $x_0=b$.

Zatem na początek należy dobrać przedział $[a,b]$ w którym funkcja $f(x)$ spełnia warunki 1)-4) oraz punkt startowy x_0 w którym spełniony jest warunek 5).

W punkcie x_0 wyznaczamy pierwszą pochodną, która reprezentuje styczną do wykresu funkcji w punkcie x_0. Styczna przecina oś x w punkcie x_1. Pochodna funkcji jest tangensem α kąta jaki tworzy styczna z osią x, stąd:

$$f'(x_0)=\tan(\alpha)=\frac{f(x_0)}{x_0-x_1}. \tag{4.9}$$

Z równania (4.9) wyznaczamy pierwsze przybliżenie pierwiastka x_1:

$$x_1=x_0-\frac{f(x_0)}{f'(x_0)}. \tag{4.10}$$

Równanie (4.10) formułuje wzór iteracyjny metody Newtona-Raphsona w postaci:

$$x_{k+1}=x_k-\frac{f(x_k)}{f'(x_k)} \text{ dla } k=0,1,2.... \tag{4.11}$$

Kryterium zakończenia iteracji stanowi najczęściej warunek $f(x_k)<\varepsilon$ lub/i $|x_{k+1}-x_k|<\varepsilon$, gdzie ε jest zadanym błędem obliczeń.

Metodę Newtona-Raphsona charakteryzują następujące cechy:

- wykładnik zbieżności jest równy $p=2$, co oznacza, że metoda jest lokalnie zbieżna kwadratowo,
- stała C jest równa $C=-0.5\frac{f''(\bar{x})}{f'(\bar{x})}$, patrz równanie (4.2),
- jest najszybszą metodą spośród prezentowanych,
- wymaga jawnej znajomości pochodnych funkcji,
- ze względu na warunki zbieżności ma mały tzw. promień zbieżności co w praktyce oznacza, że przedział izolacji $[a,b]$ nie może być duży.

Zadanie 4.1. Zaimplementuj metody bisekcji, iteracji prostej, Newtona-Raphsona z użyciem języka Python i wykorzystaj zaimplementowane procedury do rozwiązania zadań:

a) wyznacz pierwiastek funkcji $f(x)=e^{2x}-1+x$ w przedziale $[-1,2]$,

b) rozwiąż równanie $k^5-3k^4+2k^3=-2k^2+3k-1$.

5. Aproksymacja i interpolacja funkcji

5.1. Aproksymacja funkcji

Zadanie aproksymacji ciągłej polega na wyznaczeniu w danym przedziale $[a,b]$ funkcji $F(x)$, która w pewnym sensie najbardziej przybliża funkcję $f(x)$. Przyjęcie odpowiedniej klasy funkcji $F(x)$ i warunku szacowania błędu aproksymacji prowadzi do różnych sformułowań. Normę z różnicy wartości funkcji $f(x)$ i $F(x)$ zapisujemy jako: $\|f(x)-F(x)\|$.

W przypadku zadania aproksymacji dyskretnej zmienia się postać funkcji F z ciągłej $F(x)$ na dyskretną $F(x_i)$ dla $i=0,1\ldots n$, jak również norma liczona jest w postaci dyskretnej jako różnica wartości funkcji f i F w węzłach aproksymacji $i=0,1\ldots n$: $\|f(x_i)-F(x_i)\|$.

W niniejszych materiałach ograniczono się do aproksymacji dyskretnej funkcji jednej zmiennej, z błędem aproksymacji szacowanym za pomocą normy średnio-kwadratowej:

$$\|f(x_i)-F(x_i)\|=\sqrt{\sum_{i=1}^{n}\left(f(x_i)-F(x_i)\right)^2}. \tag{5.1}$$

Wyznaczenie funkcji aproksymującej $F(x)$ wymaga skorzystania z metody najmniejszych kwadratów, w której błąd aproksymacji ε:

$$\varepsilon=\sum_{i=1}^{n}\left(f(x_i)-F(x_i)\right)^2 \tag{5.2}$$

podlega minimalizacji.

Pierwszym etapem aproksymacji jest wybór funkcji bazowych aproksymacji. Funkcje te powinny być liniowo niezależne np. $[x,\sin(x),\cos(x)]$ (przykład funkcji liniowo zależnych: $[x,2x]$ lub $[\sin(x),3\sin(x)]$. Funkcje bazowe możemy przedstawić w postaci wektora o długości $m+1$:

$$\boldsymbol{\varphi}(x)=\begin{bmatrix}\varphi_0(x)\\ \varphi_1(x)\\ \vdots\\ \varphi_m(x)\end{bmatrix}. \tag{5.3}$$

Należy pamiętać aby $m<n$.

Funkcja aproksymująca stanowi kombinację liniową funkcji bazowych:

$$F(x)=\sum_{j=0}^{m}a_j\varphi_j(x). \tag{5.4}$$

Wzór (5.4) można przedstawić w formie wektorowej:

$$F(x)=\boldsymbol{a}\cdot\boldsymbol{\varphi}(x) \tag{5.5}$$

gdzie $\boldsymbol{a}$ jest wektorem współczynników aproksymacji podlegającym wyznaczeniu:

$$\boldsymbol{a}=\begin{bmatrix}a_0 & a_1 & \dots & a_m\end{bmatrix}. \tag{5.6}$$

W kolejnym etapie wprowadzamy do równania na błąd kwadratowy (5.2) przyjętą postać funkcji aproksymującej (5.5) co prowadzi do wyrażenia:

$$\varepsilon(\boldsymbol{a})=\sum_{i=1}^{n}\left(f(x_i)-\boldsymbol{a}\cdot\boldsymbol{\varphi}(x_i)\right)^2 \tag{5.7}$$

Wartość minimalną błędu (5.7) wyznaczamy z warunków zerowania się pochodnych cząstkowych liczonych dla wszystkich współczynników a_j (warunek minimum funkcji):

$$\frac{\partial\varepsilon(a_0,a_1\dots a_m)}{\partial a_j}=0\ \text{ dla }\ j=0,1\dots m \tag{5.8}$$

Warunki (5.8) tworzą liniowy układ $m+1$ równań z niewiadomymi wartościami współczynników a_j podlegający rozwiązaniu.

Podstawiając do równania (5.8) równanie (5.7) otrzymujemy:

$$\frac{\partial}{\partial a_j}\sum_{i=1}^{n}\left(f(x_i)-\boldsymbol{a}\cdot\boldsymbol{\varphi}(x_i)\right)^2=0 \text{ dla } j=0,1\ldots m. \tag{5.9}$$

Aby wyznaczyć pierwsze równanie ($k=0$) z układu równań (5.9) należy policzyć pochodną cząstkową po zmiennej a_0 korzystając z przepisu na pochodną złożoną:

$$\frac{\partial}{\partial a_0}\sum_{i=1}^{n}\left(f(x_i)-\boldsymbol{a}\cdot\boldsymbol{\varphi}(x_i)\right)^2=2\cdot\sum_{i=1}^{n}\left(f(x_i)-\boldsymbol{a}\cdot\boldsymbol{\varphi}(x_i)\right)\cdot\frac{\partial}{\partial a_0}\sum_{i=1}^{n}\left(-\boldsymbol{a}\cdot\boldsymbol{\varphi}(x_i)\right)=0 \tag{5.10}$$

gdzie $\frac{\partial}{\partial a_0}\sum_{i=1}^{n}\left(-\boldsymbol{a}\cdot\boldsymbol{\varphi}(x_i)\right)=-\sum_{i=1}^{n}\varphi_0(x_i)$ stąd otrzymujemy równanie:

$$-2\cdot\sum_{i=1}^{n}\left(f(x_i)-\boldsymbol{a}\cdot\boldsymbol{\varphi}(x_i)\right)\cdot\sum_{i=1}^{n}\varphi_0(x_i)=0. \tag{5.11}$$

Równanie (5.11) możemy przekształcić do postaci:

$$\sum_{i=1}^{n}\left(\boldsymbol{a}\cdot\boldsymbol{\varphi}(x_i)\cdot\varphi_0(x_i)\right)=\sum_{i=1}^{n}\varphi_0(x_i)\cdot f(x_i) \tag{5.12}$$

lub w wersji jawnej:

$$\sum_{i=0}^{n}\left(\sum_{j=0}^{m}a_j\cdot\varphi_j(x_i)\cdot\varphi_0(x_i)\right)=\sum_{i=1}^{n}\varphi_0(x_i)\cdot f(x_i). \tag{5.13}$$

Wszystkie pozostałe równania wyznaczane są w taki sam sposób, licząc pochodne cząstkowe po kolejnych współczynnikach a_j dla $j=1,2\ldots m$. Ogólną postać układu równań dla $k=0,1,2\ldots m$ przedstawiają równania:

$$\sum_{i=1}^{n}\left(\boldsymbol{\varphi}^{\boldsymbol{T}}(x_i)\cdot\boldsymbol{\varphi}(x_i)\right)\cdot\boldsymbol{a}=\sum_{i=1}^{n}\boldsymbol{\varphi}(x_i)\cdot f(x_i) \tag{5.14}$$

lub

$$\sum_{i=0}^{n}\left(\sum_{j=0}^{m}a_j\cdot\varphi_j(x_i)\cdot\varphi_k(x_i)\right)=\sum_{i=1}^{n}\varphi_k(x_i)\cdot f(x_i) \text{ dla } k=0,1,2\ldots m. \tag{5.15}$$

Wyodrębniając z równania (5.14) macierze współczynników układu równań, nazwijmy ją $\boldsymbol{A}$, i wektora prawej strony, nazwijmy go $\boldsymbol{b}$ otrzymujemy:

$$\boldsymbol{A}=\sum_{i=1}^{n}\left(\boldsymbol{\varphi}^{\boldsymbol{T}}(x_i)\cdot\boldsymbol{\varphi}(x_i)\right),\ \boldsymbol{b}=\sum_{i=1}^{n}\boldsymbol{\varphi}(x_i)\cdot f(x_i), \tag{5.16}$$

Tak samo możemy przedstawić wersję równania (5.15):

$$A_{j,k}=\sum_{i=1}^{n}\left(\varphi_j(x_i)\cdot\varphi_k(x_i)\right),\ b_j=\sum_{i=1}^{n}\varphi_j(x_i)\cdot f(x_i) \text{ dla } j,k=0,1\ldots m \tag{5.17}$$

Ogólna postać układu równań aproksymacji przyjmuje formę:

$$\boldsymbol{A}\cdot\boldsymbol{a}=\boldsymbol{b}, \tag{5.17}$$

której rozwiązaniem jest wektor współczynników aproksymacji $\boldsymbol{a}$. Stąd otrzymujemy funkcję aproksymacyjną w postaci $F(x)=\sum_{j=0}^{m} a_j\varphi_j(x)$ lub $F(x)=\boldsymbol{a}\cdot\boldsymbol{\varphi}(x)$.

Jeżeli wektor funkcji bazowych stanowią jednomiany np. $[1 \quad x \quad x^2 \quad x^3]$ to wówczas mówimy o aproksymacji wielomianowej. Aproksymacja wielomianowa jest najczęściej spotykaną aproksymacją w praktycznych zagadnieniach.

Przykłady aproksymacji:

- Aproksymacja liniowa to aproksymacja wielomianowa z dwoma funkcjami bazowymi $\boldsymbol{\varphi}(x)=[1 \quad x]$.
- Aproksymacja kwadratowa to aproksymacja wielomianowa z trzema funkcjami bazowymi $\boldsymbol{\varphi}(x)=[1 \quad x \quad x^2]$.
- Aproksymacja trygonometryczna $\boldsymbol{\varphi}(x)=[1 \quad \sin(x) \quad \cos(x) \quad \sin(2x) \quad \cos(2x)]$.
- Aproksymacja wykładnicza $\varphi(x)=a_0 \mathrm{e}^{a_1 x}$.
- Aproksymacja potęgowa $\varphi(x)=a_0 \mathrm{x}^{a_1}$.

Uwaga

Dwa ostatnie przykłady prowadzą do nieliniowych układów równań. Jednak, poprzez wprowadzenie nowych zmiennych $X_i=\ln(x_i)$, $Y_i=\ln(f(x_i))$, $A_0=\ln(a_0)$, $A_1=a_1$ możemy zamienić te problemy na problem aproksymacji liniowej z funkcją aproksymującą w postaci: A_0+A_1X, co zamienia układ równań na liniowy.

5.2. Interpolacja funkcji

Interpolacja jest szczególnym przypadkiem aproksymacji dyskretnej, kiedy błąd aproksymacji $\varepsilon=0$, może mieć to miejsce wówczas kiedy funkcja aproksymująca $F(x)$ przechodzi przez wszystkie węzły aproksymacji w badanym przedziale $[a,b]$. Takie zadanie aproksymacji jest możliwe do rozwiązania gdy liczba funkcji bazowych jest równa liczbie węzłów aproksymacji tzn. $m=n$. Z tego wynika, że algorytm budowania funkcji aproksymacyjnej może być zastosowany do rozwiązania zadania interpolacji.

Jednak model zadania interpolacji, czyli układ równań z niewiadomymi współczynnikami interpolacji możemy wyprowadzić w prostszy sposób wychodząc z warunku interpolacji w postaci:

$$F(x_i)=f(x_i) \text{ dla } i=0,1\ldots n, \tag{5.18}$$

co oznacza, że wartość funkcji interpolującej $F(x_i)$ w węzłach interpolacji $i=0,1\ldots n$ jest dokładnie równa wartości zadanej $f(x_i)$.

Podstawiając do równania (5.18) aproksymację w postaci (5.4) otrzymujemy układ równań:

$$\sum_{j=0}^{m} a_j \varphi_j(x_i)=f(x_i) \text{ dla } i=0,1\ldots n \text{ i } n=m \tag{5.19}$$

z niewiadomymi współczynnikami interpolacji a_j.

To samo równanie możemy zapisać w formie macierzowej:

$$\boldsymbol{a}\cdot\boldsymbol{\varphi}(x_i)=f(x_i) \text{ dla } i=0,1\ldots n \tag{5.20}$$

Podobnie jak w przypadku aproksymacji, możemy mówić o interpolacji liniowej, kwadratowej lub ogólnie wielomianowej jeżeli funkcjami bazowymi są jednomiany i jest to również najczęściej spotykana w praktyce interpolacja.

Rozwiązaniem zadania jest rozwiązanie układu równań (5.19) lub (5.20), jednakże w przypadku interpolacji wielomianowej, możemy się posłużyć jawnymi wzorami tzw. interpolacji Lagrange'a.

Interpolacja Lagrange'a

W ogólnym przypadku, aby skonstruować wielomian interpolacyjny Lagrange'a stopnia m, należy zdefiniować zbiór $m+1$ wielomianów bazowych Lagrange'a m-tego stopnia dla każdego węzła interpolacji $i=0,1\ldots n$. Dla węzła i wielomian bazowy Lagrange'a stopnia m jest konstruowany jako:

$$L_{m,i}(x)=\prod_{\substack{i=0\\ j\neq i}}^{n} \frac{x-x_j}{x_i-x_j} \text{ dla } n=m, \tag{5.21}$$

lub w formie rozwiniętej:

$$L_{m,i}(x)=\frac{(x-x_0)\ldots(x-x_{i-1})(x-x_{i+1})\ldots(x-x_n)}{(x_i-x_0)\ldots(x_i-x_{i-1})(x_i-x_{i+1})\ldots(x_i-x_n)}. \tag{5.22}$$

Wielomiany bazowe Lagrange'a $L_{m,i}(x)$ przecinają oś x w każdym węźle interpolacji, z wyjątkiem węzła, dla którego ten wielomian został zbudowany, co jest wyrażone przez własność Kroneckera:

$$\sum_{i=0}^{m} L_{m,i}(x)=1. \tag{5.24}$$

Zatem wielomian interpolacyjny Lagrange'a m-tego stopnia ma postać:

$$F_m(x)=\sum_{i=0}^{m} f(x_i)L_{m,i}(x). \tag{5.25}$$

Można łatwo zauważyć, że dla dowolnego węzła x_i wartość wielomianu Lagrange'a jest równa $f(x_i)$ co spełnia założenie interpolacji.

Interpolacja Hermite'a

Interpolacja Hermite'a pozwala na wyznaczenie wielomianu zgodnego w węzłach interpolacji nie tylko z wartościami danej funkcji ale także z jej pochodnymi. Te dodatkowe warunki powodują podniesienie stopnia wielomianu interpolacyjnego Hermite'a do rzędu $2m+1$. Postać wielomianu Hermite'a jest dana wzorem:

$$H_{2m+1}(x)=\sum_{i=0}^{m} f(x_i)H_{m,i}(x)+\sum_{i=0}^{m} f'(x_i)\widehat{H}_{m,i}(x). \tag{5.26}$$

W powyższym wzorze pierwszy człon, po prawej stronie, jest powiązany z wartościami funkcji, natomiast drugi człon uwzględnia pochodne danej funkcji. Funkcje bazowe dla pierwszego członu mają postać:

$$H_{m,i}(x)=[1-2(x-x_i)L'_{m,i}(x_i)]L^2_{m,i}(x). \tag{5.27}$$

W drugim członie funkcje bazowe są definiowane jako:

$$\widehat{H}_{m,i}(x)=(x-x_i)L^2_{m,i}(x). \tag{5.28}$$

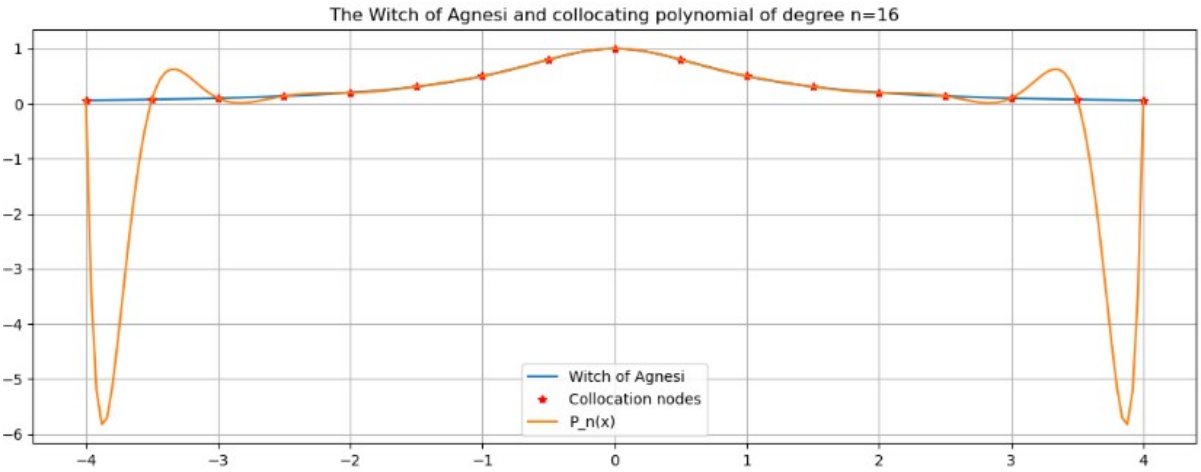
gdzie $L_{m,i}(x)$ jest wielomianem bazowym Lagrange'a dla węzła i stopnia m, oraz $L'_{m,i}(x_i)$ jest wartością pierwszej pochodnej wielomianu bazowego Lagrange'a w węźle x_i.

Uwaga

Wysokie stopnie wielomianów interpolacyjnych dla równo-odlegle rozłożonych węzłów nie są zalecane, ze względu na oscylacje wielomianu na krańcach przedziałów co wprowadza znaczący błąd interpolacji mimo spełnienia warunku zerowania błędu w węzłach interpolacji. Rysunek 1 przedstawia interpolację wielomianem stopnia $n=16$ z 17-ma równo-oddalonymi węzłami w przedziale $[-4,4]$

funkcji $\frac{1}{1+x^2}$ (nazywanej *Witch of Agnesi*). Na końcach przedziałów zaobserwować można bardzo duże odchylenia interpolacji od krzywej interpolowanej. Rozwiązaniem tego problemu może być użycie interpolacji funkcjami sklejanymi z wielomianów niskiego rzędu lub aproksymacja wielomianem niższego rzędu.

[Tekst alternatywny. Przedruk wykresu wielomianu interpolacyjnego stopnia 16 z pozycji: Brenton LeMesurier „Introduction to Numerical Methods and Analysis with Python" str. 152]



Rys. 1. Interpolacja wielomianem stopnia n=16.

Zadanie 5.1. Rozwiąż zadanie aproksymacji liniowej dla danych

x_i	1	2	3	4	5	6	7	8	9	10
$f(x_i)$	1.3	3.5	4.2	5.0	7.0	8.8	10.1	12.5	13.0	15.6

Zadanie 5.2. Rozwiąż zadanie aproksymacji potęgowej dla danych z zadania 5.1.

Zadanie 5.3. Rozwiąż zadanie interpolacji Hermite'a dla następujących danych:

x_i	$f(x_i)$	$f'(x_i)$
8.3	17.56	1.11
8.6	18.50	1.15

6. Numeryczne całkowanie

Głównym celem metod całkowania numerycznego (zwanych również kwadraturami) jest wyznaczanie całek, których obliczenie analityczne jest niemożliwe lub bardzo trudne. Zakres tego materiału ogranicza się do całek pojedynczych, co oznacza, że funkcja podcałkowa jest funkcją jednej zmiennej niezależnej (funkcja podcałkowa to funkcja, dla której oblicza się całkę). Rozważany problem polega na oszacowaniu wartości następującej całki oznaczonej:

$$I=\int_a^b f(x)dx, \tag{6.1}$$

gdzie a i b stanowią granice całkowania.

Analityczne wyznaczenie całki (6.1) wymaga znajomości funkcji pierwotnej $F(x)$ dla funkcji podcałkowej $f(x)$, wówczas całka oznaczona jest wyrażona jako różnica wartości funkcji pierwotnej w granicach całkowania i jest to wówczas rozwiązanie dokładne $I=F(b)-F(a)$. Kiedy funkcja pierwotna nie jest znana i nie może być znaleziona w formie jawnej jedyną metodą jest wówczas wyznaczenie całki z użyciem technik numerycznych. Innym przykładem kiedy konieczne jest użycie technik numerycznych jest problem obliczenia całki dla funkcji podcałkowej zadanej w sposób dyskretny np. obliczenie pola przekroju zbiornika wodnego na danym odcinku, gdzie pomierzona została głębokość tylko w określonych punktach.

W najprostszej postaci rozwiązaniem jest interpolacja wielomianowa funkcji podcałkowej. Dla tak skonstruowanego przybliżenia, które jest funkcją wielomianową określonego stopnia, łatwo jest znaleźć funkcję pierwotną. Takie podejście prowadzi do tzw. kwadratur Newtona-Cotesa.

6.1. Kwadratury Newtona-Cotesa

Kwadratura Newtona-Cotesa jest to kwadratura $Q(f)=I(L_n)$, gdzie L_n jest wielomianem interpolacyjnym Lagrange'a opartym na $n+1$ równoodległych węzłach. Zatem w celu zbudowania kwadratury Newtona-Cotesa w przedziale $[a,b]$ należy wyznaczyć równoodległe węzły $x_0, x_1 \ldots x_n$ (gdzie $x_0=a$ i $x_n=b$), czyli:

$$x_k=x_0+k\frac{b-a}{n} \text{ dla } k=0,1\ldots n. \tag{6.2}$$

W całce (6.1) zastępujemy funkcję $f(x)$ wielomianem interpolacyjnym Lagrange'a stopnia n:

$$I=\int_a^b f(x)dx \approx \int_a^b L_n(x)dx, \tag{6.3}$$

$$L_n(x)=\sum_{k=0}^{n} L_{n,k}(x)f(x_k) \tag{6.4}$$

gdzie $L_{n,k}(x)$ jest bazowym wielomianem interpolacyjnym Lagrange'a stopnia n dla węzła k wyrażonym równaniem w formie (5.22).

W zależności od przyjętego stopnia interpolacji otrzymujemy metody:

- metoda prostokątów dla $n=0$, liczba węzłów interpolacji 1 ($x_0=a$ lub $x_0=b$),
- metoda trapezów dla $n=1$, liczba węzłów interpolacji 2 ($x_0=a$, $x_1=b$),
- metoda Simpsona dla $n=2$, liczba węzłów interpolacji 3 ($x_0=a$, $x_1=\frac{a+b}{2}$, $x_2=b$).

Nie stosuje się praktycznie metod wyższego rzędu ze względu na problem przedstawiony w uwagach do punktu 5.2. W przypadkach kiedy w obszarze całkowania funkcja interpolująca niedostatecznie przybliża funkcję podcałkową, wprowadzając zbyt duży błąd obliczeń, możemy zastosować kwadratury złożone. Polegają one na podziale obszaru na N pod-obszarów o równej długości, wyznaczeniu w tych pod-obszarach za pomocą kwadratur prostych całek i zsumowaniu tych wartości.

Kwadratury proste

- **Metoda prostokątów**

Metoda prostokątów prosta jest policzeniem całki jako pola prostokąta o wymiarze $f(a)\cdot(b-a)$ lub $f(b)\cdot(b-a)$, lub korzystając z ogólnego zapisu kwadratury Newtona-Cotesa (6.3), dla $L_0(x)=f(a)$ lub $f(b)$:

$$I=\int_a^b f(x)dx\approx\int_a^b L_0(x)dx=\int_a^b f(a)dx=f(a)\cdot(b-a), \tag{6.5}$$

lub

$$I=\int_a^b f(x)dx\approx\int_a^b L_0(x)dx=\int_a^b f(b)dx=f(b)\cdot(b-a). \tag{6.6}$$

Odmianą metody prostokątów jest metoda punktu środkowego, która wykorzystuje do obliczenia pola prostokąta wartość funkcji w punkcie środkowym przedziału $[a,b]$ czyli:

$$I=\int_a^b f(x)dx\approx\int_a^b f(\frac{a+b}{2})dx=f(\frac{a+b}{2})\cdot(b-a). \tag{6.7}$$

- **Metoda trapezów**

W metodzie trapezów funkcja podcałkowa przybliżana jest wzorem interpolacyjnym 1 -stopnia, zatem $n=1$, $x_0=a$ i $x_1=b$:

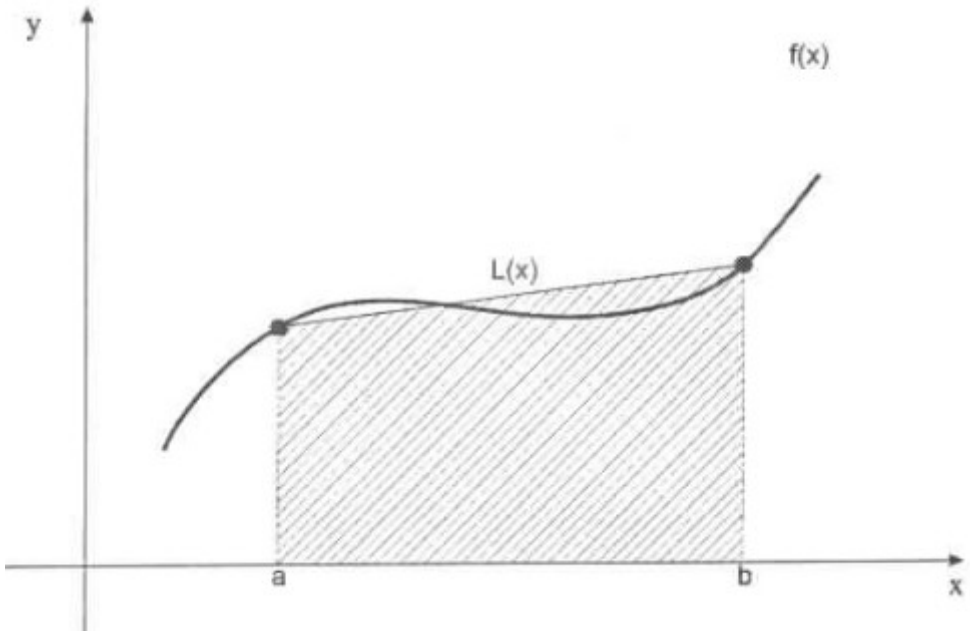
$$I=\int_a^b f(x)dx\approx\int_a^b L_1(x)dx=\int_a^b\left(f(a)\frac{x-x_1}{x_0-x_1}+f(b)\frac{x-x_0}{x_1-x_0}\right)dx=\frac{x_1-x_0}{2}\cdot(f(x_0)+f(x_1)) \qquad (6.8)$$

Rysunek 2 przedstawia graficzną interpretację wzoru trapezów.

Zatem kwadraturę trapezów dla jednego przedziału $[a,b]$ możemy wyrazić jako:

$$Q(f)=\frac{b-a}{2}\cdot(f(a)+f(b)). \qquad (6.9)$$

[Tekst alternatywny. Przedruk rysunku graficznej interpretacji wzoru trapezów z pozycji: Ewa Dudek-Dyduch, Jarosław Wąs, Lidia Dudkiewicz, Katarzyna Grobler-Dębska, Bratłomiej Gudowski „Metody numeryczne. Wybrane zagadnienia”, rys. 6.1, str. 87]



Rys. 2. Graficzna interpretacja wzoru trapezów.

- **Metoda Simpsona**

Postępując tak samo jak w poprzednich przypadkach kwadratur, tym razem interpolując funkcję podcałkową wielomianem interpolacyjnym Lagrange’a stopnia $n=2$ czyli $L_2(x)$, wykonując jawne całkowanie wielomianu drugiego stopnia w przedziale $[a,b]$ otrzymujemy wzór Simpsona w postaci:

$$Q(f)=\frac{x_2-x_0}{6}\cdot(f(x_0)+4\cdot f(x_1)+f(x_2)), \tag{6.10}$$

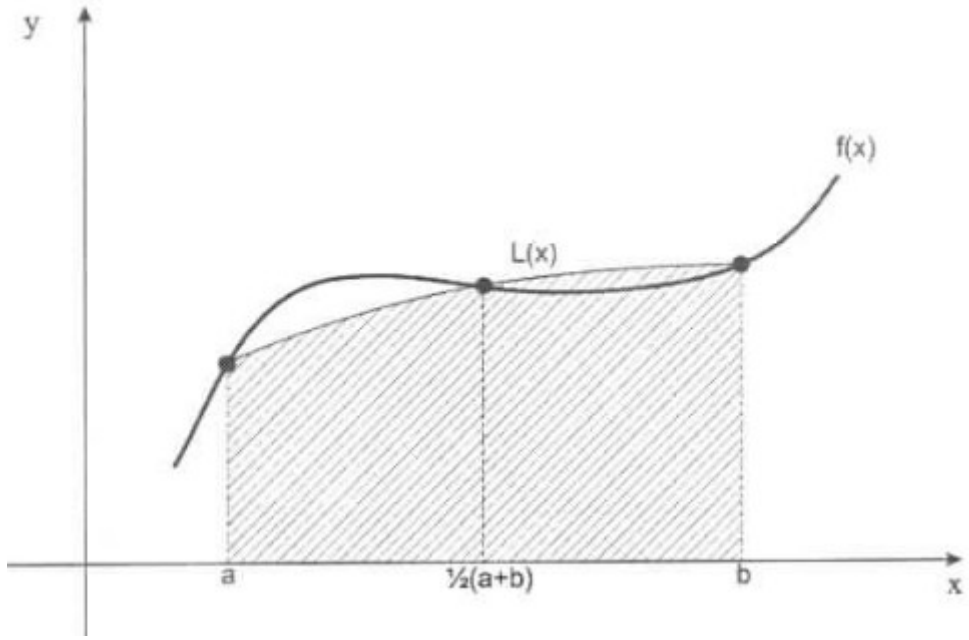
gdzie $x_0=a$, $x_1=\frac{a+b}{2}$ i $x_2=b$.

Wzór możemy zapisać w równoważnej postaci z wykorzystaniem wartości granic przedziału całkowania:

$$Q(f)=\frac{b-a}{6}\cdot(f(a)+4\cdot f(\frac{a+b}{2})+f(b)). \tag{6.11}$$

Graficzną interpretację wzoru Simpsona przedstawia rys. 2.

[Tekst alternatywny. Przedruk rysunku graficznej interpretacji wzoru trapezów z pozycji: Ewa Dudek-Dyduch, Jarosław Wąs, Lidia Dudkiewicz, Katarzyna Grobler-Dębska, Bratłomiej Gudowski „Metody numeryczne. Wybrane zagadnienia”, rys. 6.2, str. 88]



Rys. 3. Graficzna interpretacja wzoru Simpsona.

Kwadratury złożone

W kwadraturach złożonych Newtona-Cotesa wybrany przedział całkowania dzielimy na N pod-przedziałów z równoodległymi węzłami:

$$x_i=a+i\cdot\frac{b-a}{N} \text{ dla } i=0,1\ldots N. \tag{6.12}$$

Następnie w każdym pod-przedziale należy zastosować wybraną kwadraturę Newtona-Cotesa i wyniki zsumować:

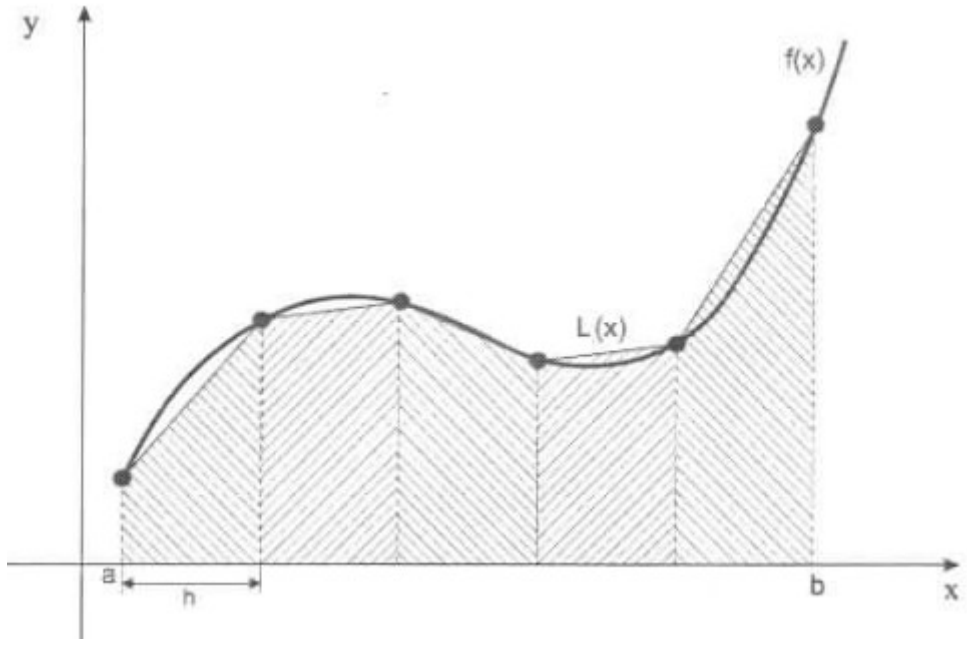
$$I=\int_a^b f(x)\,dx \approx \sum_{i=1}^{N}\left(\int_{x_{i-1}}^{x_i} L_n(x)\,dx\right) \tag{6.13}$$

Kwadratura (6.13) zastosowana do wzoru trapezów prowadzi do **złożonej kwadratury trapezów** w postaci:

$$Q_N(f)=\sum_{i=1}^{N}\left(\frac{x_i-x_{i-1}}{2}\cdot\left(f(x_{i-1})+f(x_i)\right)\right) \tag{6.14}$$

Rysunek 3 przedstawia graficzną interpretację złożonej metody trapezów.

[Tekst alternatywny. Przedruk rysunku graficznej interpretacji złożonej kwadratury trapezów z pozycji: Ewa Dudek-Dyduch, Jarosław Wąs, Lidia Dudkiewicz, Katarzyna Grobler-Dębska, Bratłomiej Gudowski „Metody numeryczne. Wybrane zagadnienia”, rys. 6.3, str. 90]



Rys. 4. Graficzna interpretacja złożonej metody trapezów.

Zastosowanie interpolacji Lagrange’a rzędu drugiego we wzorze (6.13) prowadzi do **złożonej kwadratury Simpsona** w formie:

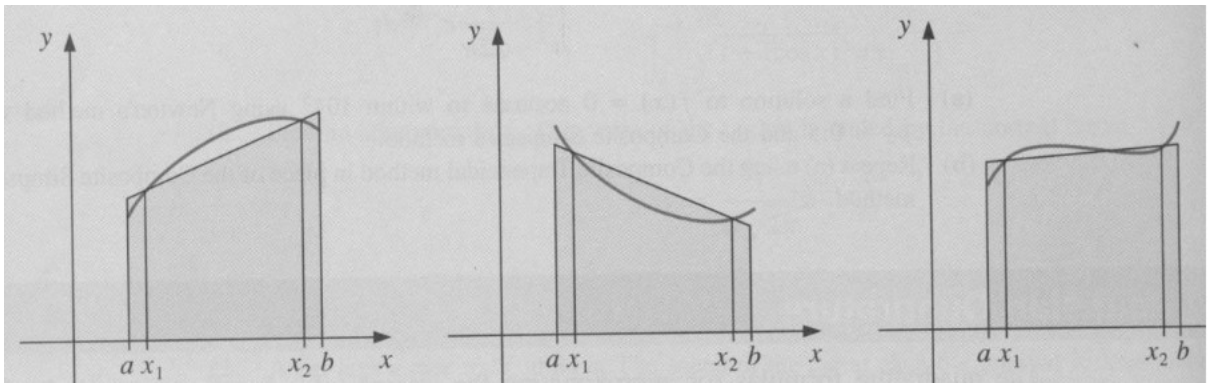
$$Q_N(f)=\sum_{i=1}^{N}\left(\frac{x_i-x_{i-1}}{6}\cdot\left(f(x_{i-1})+4f\left(\frac{x_{i-1}+x_i}{2}\right)+f(x_i)\right)\right), \tag{6.15}$$

dla wartości współrzędnych węzłów aproksymacji obliczonych z wzoru (6.12).

6.2. Kwadratury Gaussa

Kwadratury Newtona-Cotesa zakładają, że funkcja interpolująca przechodzi przez węzły zdefiniowane przez krańce przedziałów. Nie jest to jednak najlepsze przybliżenie funkcji. W większości przypadków interpolacja taka jak pokazana na rysunku 5 dałaby lepsze przybliżenie przy tym samym stopniu niż interpolacja przechodząca przez węzły a, b. Kwadratura Gaussa szacuje współrzędne węzłów w sposób optymalny, a nie w równych odstępach, pozwalając na lepszy wynik w porównaniu do metod Newtona-Cotes.

[Tekst alternatywny. Przedruk rysunku interpolacji z węzłami wewnątrz obszaru $[a,b]$ z pozycji: Terrence J. Akai, „Applied numerical methods for engineers": John Wiley & Sons, cop. 1994.]



Rys. 5. Interpolacja z węzłami wewnątrz obszaru $[a,b]$.

Kwadratura Gaussa korzysta z optymalnych położeń węzłów x_i i współczynników interpolacji c_i. Aby uprościć obliczenia tych nieznanych wartości rozważana jest całka w znormalizowanym przedziale $t\in[-1,1]$ wyrażona przez:

$$I=\int_{-1}^{1}\check{f}(t)\,dt, \tag{6.16}$$

gdzie t jest znormalizowanym układem współrzędnych. Całka (6.16) wyrażana jest przez kombinację liniową współczynników interpolacji c_i i wartości funkcji w węzłach interpolacji $\check{f}(t_i)$:

$$\int_{-1}^{1}\check{f}(t)\,dt=\sum_{i=0}^{n-1}c_i\cdot\check{f}(t_i). \tag{6.17}$$

Rozważmy rozwiązania dokładne z dwoma węzłami interpolacji $n=1$ (co oznacza interpolację liniową) dla jednomianów $1,t,t^2,t^3$:

$$\begin{cases} I(\check{f}(t)=1)=\int_{-1}^{1}(1)\,dt=1-(-1)=2=\sum_{i=0}^{1}c_i\cdot\check{f}(t_i)=c_0+c_1 \\ I(\check{f}(t)=t)=\int_{-1}^{1}(t)\,dt=\frac{1}{2}-\frac{1}{2}=0=\sum_{i=0}^{1}c_i\cdot\check{f}(t_i)=c_0\cdot t_0+c_1\cdot t_1 \\ I(\check{f}(t)=t^2)=\int_{-1}^{1}(t^2)\,dt=\frac{1}{3}+\frac{1}{3}=\frac{2}{3}=\sum_{i=0}^{1}c_i\cdot\check{f}(t_i)=c_0\cdot t_0^2+c_1\cdot t_1^2 \\ I(\check{f}(t)=t^3)=\int_{-1}^{1}(t^3)\,dt=\frac{1}{4}-\frac{1}{4}=0=\sum_{i=0}^{1}c_i\cdot\check{f}(t_i)=c_0\cdot t_0^3+c_1\cdot t_1^3 \end{cases}. \tag{6.18}$$

Prowadzi to do nieliniowego układu równań ze względu na współczynniki c_0, c_1, t_0, t_1 w postaci:

$$\begin{cases} c_0+c_1=2 \\ c_0\cdot t_0+c_1\cdot t_1=0 \\ c_0\cdot t_0^2+c_1\cdot t_1^2=\frac{2}{3} \\ c_0\cdot t_0^3+c_1\cdot t_1^3=0 \end{cases}. \tag{6.19}$$

Rozwiązaniem układu równań (6.19) jest:

$$c_0=1, c_1=1, t_0=-\frac{1}{\sqrt{3}}, t_1=\frac{1}{\sqrt{3}}. \tag{6.20}$$

Postępując podobnie możemy wyznaczyć współczynniki dla interpolacji wyższych rzędów (Tabela 6.1).

Tabela 6.1. Współczynniki i współrzędne węzłów interpolacji Gaussa dla rzędów $n=1$ do 3.

n	Węzły t_i	Współczynniki c_i
1	-0.5773502692	1.0000000000
	0.5773502692	1.0000000000
2	-0.7745966692	0.5555555556
	0.0000000000	0.8888888889
	0.7745966692	0.5555555556
3	-0.8611363116	0.3478548451
	-0.3399810436	0.6521451549
	0.8611363116	0.3478548451
	0.3399810436	0.6521451549

Aby rozwiązać całkę wyjściową $I=\int_a^b f(x)\,dx$ należy dokonać transformacji współrzędnych z układu x do t, co przedstawia wzór:

$$\int_a^b f(x)dx = \int_{-1}^{1} \frac{d}{dt}x(t)\cdot f(x(t))dt \tag{6.21}$$

gdzie $\frac{d}{dt}x(t)$ jest tzw. Jacobianem przekształcenia podlegającym wyznaczeniu. Funkcję $x(t)$ możemy wyznaczyć z zależności:

$$x(t) = m\cdot t + c \tag{6.22}$$

gdzie $x=a \rightarrow t=-1, x=b \rightarrow t=1$ i $\frac{d}{dt}x(t)=m$, po rozwiązaniu proporcji otrzymujemy współczynniki przekształcenia równe $m=\frac{b-a}{2}, c=\frac{b+a}{2}$. Podstawiając transformację (6.22) do zależności (6.21) otrzymujemy:

$$\int_a^b f(x)dx = \frac{b-a}{2}\int_{-1}^{1} f\left(\frac{b-a}{2}\cdot t + \frac{b+a}{2}\right)dt. \tag{6.23}$$

Zależność (6.17) ostatecznie prowadzi do kwadratury Gaussa całki oznaczonej w przedziale $[a,b]$ stopnia interpolacji n:

$$\int_a^b f(x)dx \approx \sum_{i=0}^{n-1} c_i\cdot \check{f}(t_i) = \frac{b-a}{2}\sum_{i=0}^{n-1} c_i\cdot f\left(\frac{b-a}{2}\cdot t_i + \frac{b+a}{2}\right). \tag{6.24}$$

Tak jak w przypadku kwadratur Newtona-Cotesa możemy zastosować kwadraturę złożoną poprzez podział obszaru całkowania na N równych części, wyznaczeniu w każdym pod-obszarze całki według wzoru (6.23) i następnie ich zsumowanie zgodnie z wzorem (6.12) i (6.13). Metoda Gaussa nie posiada ogólnego wzoru kwadratury złożonej.

Zadanie 6.1. Wyznacz wartość całki wyrażonej przez:

$$\int_1^{1.5} e^{-x^2}dx,$$

dla której rozwiązanie dokładne wynosi: 0.1183197 metodami prostymi:

- trapezów,
- Simpsona,
- Gaussa

oraz wersjami złożonymi tych metod dla $N=2$. Oblicz błąd rozwiązań dla każdego przypadku.

Zadanie 6.2. Zaimplementuj złożone kwadratury z zadania 6.1 z użyciem języka programowania Python i wyznacz całkę:

$$\frac{gm}{c}\int_0^{10} 1-e^{-\frac{c}{m}x}\,dx,$$

gdzie: $g=9.8, m=68.1, c=12.5$. Rozwiązanie dokładne wynosi 289.435. Wykonaj analizę błędu w zależności od stopnia metody i liczby podziałów N.

7. Numeryczne różniczkowanie

7.1. Różnice skończone

Podstawą metod numerycznego różniczkowania jest rozwinięcie funkcji w nieskończony szereg Taylora, dla punktu x_{i+1} np. metodą wprzód:

$$f(x_{i+1})=f(x_i)+f'(x_i)h+\frac{f''(x_i)}{2}h^2+\ldots \tag{7.1}$$

gdzie $h=x_{i+1}-x_i$.

Obcinając ciąg (7.1), wykluczając drugą i wyższe pochodne otrzymamy przybliżenie pierwszej pochodnej w punkcie x_i w postaci:

$$f'(x_i)\approx\frac{f(x_{i+1})-f(x_i)}{h} \tag{7.2}$$

Rozwijając funkcję $f(x)$ w szereg Taylora dla punktu x_{i+2} otrzymujemy

$$f(x_{i+2})=f(x_i)+f'(x_i)2h+\frac{f''(x_i)}{2}(2h)^2+\ldots. \tag{7.3}$$

Pomnożenie równania (7.1) przez 2 i odjęcie go od równania (7.3) prowadzi do:

$$f(x_{i+2})-2f(x_{i+1})=f(x_i)-f''(x_i)h^2+\ldots, \tag{7.4}$$

skąd można wyznaczyć pochodną drugiego rzędu dla punktu x_i:

$$f''(x_i)\approx\frac{f(x_{i+2})-2f(x_{i+1})+f(x_i)}{h^2}. \tag{7.5}$$

Formuła (7.5) jest nazywana pochodną skończoną drugiego rzędu wprzód. Analogicznie, pochodne można wyznaczać rozwijając funkcję w szereg Taylora wstecz:

$$f(x_{i-1}) = f(x_i) - f'(x_i)h + \frac{f''(x_i)}{2}h^2 + \ldots \tag{7.6}$$

Dla metody wstecz pochodną skończoną pierwszego rzędu wyraża wzór:

$$f'(x_i) \approx \frac{f(x_i) - f(x_{i-1})}{h}, \tag{7.7}$$

a pochodną rzędu drugiego zapisujemy w postaci:

$$f''(x_i) \approx \frac{f(x_i) - 2f(x_{i-1}) + f(x_{i-2})}{h^2}. \tag{7.8}$$

Błąd obliczenia pochodnej za pomocą formuł (7.2), (7.5), (7.7) i (7.8) zmniejsza się liniowo wraz ze zmniejszaniem się odległości między węzłami x_{i-1}, x_i, x_{i+1} co zapisuje się w postaci funkcji błędu $\mathrm{O}(h)$.

Można również skorzystać z rozwinięcia centralnego funkcji w szereg Taylora. Odejmując rozwinięcie wstecz szeregu Taylora (7.6) od jego rozwinięcia wprzód (7.1) otrzymujemy rozwinięcie centralne szeregu Taylora w postaci

$$f(x_{i+1}) = f(x_{i-1}) + 2f'(x_i)h + \frac{2f^{(3)}(x_i)}{3!}h^3 + \ldots, \tag{7.9}$$

stąd skończona pochodna centralna pierwszego rzędu przyjmuje postać:

$$f'(x_i) \approx \frac{f(x_{i-1}) - f(x_{i+1})}{2h}. \tag{7.10}$$

Natomiast pochodną skończoną drugiego rzędu wyraża wzór:

$$f''(x_i) \approx \frac{f(x_{i-1}) - 2f(x_i) + f(x_{i+1})}{h^2}. \tag{7.11}$$

W przypadku formuł centralnych (7.10), (7.11) funkcja błędu pochodnej jest proporcjonalna do kwadratu odległości między węzłami co wyraża się przez $\mathrm{O}(h^2)$.

Postępując podobnie można wyznaczać pochodne wyższych rzędów, jednak ten problem nie jest rozważany w niniejszych materiałach.

7.2. Różnice skończone wysokiej dokładności

Zwiększenie dokładności formuł różnic skończonych może być osiągnięte poprzez uwzględnienie dodatkowych członów w rozwinięciu funkcji w szereg Taylora. Tym razem do wyznaczenia pochodnej skończonej wprzód posłużymy się rozwinięciem Taylora (7.1) z uwzględnieniem trzeciego członu, za który podstawiona zostanie pochodna skończona drugiego rzędu wprzód (7.5) co prowadzi do:

$$f'(x_i) \approx \frac{f(x_{i-1}) - f(x_i)}{h} - \frac{f(x_{i+2}) - 2f(x_{i+1}) + f(x_i)}{2h} \tag{7.12}$$

lub w formie uporządkowanej:

$$f'(x_i) \approx \frac{-f(x_{i+2}) + 4f(x_{i+1}) - 3f(x_i)}{2h}. \tag{7.13}$$

Dla pochodnej drugiego rzędu otrzymujemy skończoną różnicę w postaci:

$$f''(x_i) \approx \frac{-f(x_{i+3}) + 4f(x_{i+2}) - 5f(x_{i+1}) + 2f(x_i)}{h^2}. \tag{7.14}$$

Należy zauważyć, że uwzględnienie członu drugiej pochodnej poprawia dokładność do rzędu $O(h^2)$.

Podobne, ulepszone wersje można opracować dla wzorów wstecz i centralnych, a także dla aproksymacji wyższych pochodnych. Wzory dla pochodnej rzędu pierwszego i drugiego dla różnic skończonych wprzód, wstecz oraz centralnych przedstawia Tabela 7.1. W tabeli 7.1 podano również rząd zbieżności w postaci funkcji błędu.

Tabela 7.1. Wzory skończenie różnicowe dla pochodnych pierwszego $f'(x)$ i drugiego rzędu $f''(x)$.

Wzór	$f'(x)$	$f''(x_i)$	*Błąd*
wprzód	$\frac{f(x_{i+1}) - f(x_i)}{h}$	$\frac{f(x_{i+2}) - 2f(x_{i+1}) + f(x_i)}{h^2}$	$O(h)$
wprzód dokładny	$\frac{-f(x_{i+2}) + 4f(x_{i+1}) - 3f(x_i)}{2h}$	$\frac{-f(x_{i+3}) + 4f(x_{i+2}) - 5f(x_{i+1}) + 2f(x_i)}{h^2}$	$O(h^2)$
wstecz	$\frac{f(x_i) - f(x_{i-1})}{h}$	$\frac{f(x_i) - 2f(x_{i-1}) + f(x_{i-2})}{h^2}$	$O(h)$
wstecz dokładny	$\frac{3f(x_i) - 4f(x_{i-1}) + f(x_{i-2})}{2h}$	$\frac{2f(x_i) - 5f(x_{i-1}) + 4f(x_{i-2}) - f(x_i - 3)}{h^2}$	$O(h^2)$
centralny	$\frac{f(x_{i-1}) - f(x_{i+1})}{2h}$	$\frac{f(x_{i-1}) - 2f(x_i) + f(x_{i+1})}{h^2}$	$O(h^2)$
centralny dokładny	$\frac{-f(x_{i+2}) + 8f(x_{i+1}) + f(x_{i-1}) - 8f(x_{i-2})}{12h}$		$O(h^4)$
	$\frac{-f(x_{i+2}) + 16f(x_{i+1}) - 30f(x_i) + 16f(x_{i-1}) - f(x_{i-2})}{12h^2}$		$O(h^4)$

Zadanie 7.1. Wyznacz pierwszą pochodną funkcji:

$$f(x)=-0.1x^4-0.15x^3-0.5x^2-0.25x+1.2$$

w punkcie $x=0.5$ z krokiem $h=0.25$ używając wzorów skończenie różnicowych z Tabeli 7.1. Policz błąd oszacowania porównując wyniki z rozwiązaniem dokładnym -0.9125.

Zadanie 7.2. Oblicz przybliżenia pochodnych wzorami różnic skończonych metodami wprzód, wstecz i centralną dla pierwszej i drugiej pochodnej $f(x)=\cos(x)$ w punkcie $x=\frac{\pi}{4}$, używając $h=\frac{\pi}{12}$. Oszacuj procentowy błąd względny dla każdego przybliżenia.

8. Numeryczne metody rozwiązywania równań różniczkowych

Równania, które składają się z nieznanej funkcji i jej pochodnych, nazywane są *równaniami różniczkowymi* np. poniższe równanie oparte na drugim prawie Newtona pozwala obliczyć prędkość v spadającego obiektu w funkcji czasu t:

$$\frac{dv}{dt}=g-\frac{c}{m}v, \tag{8.1}$$

gdzie g jest stałą grawitacji, c jest współczynnikiem oporu powietrza, m jest masą opadającego obiektu i t jest zmienną czasu. W równaniu (8.1) różniczkowana wielkość, v, nazywana jest zmienną zależną. Wielkość, względem której różniczkujemy v, czyli t, nazywana jest zmienną niezależną. Gdy funkcja obejmuje jedną zmienną niezależną, równanie nazywa się *równaniem różniczkowym zwyczajnym* (lub ODE). Jest to przeciwieństwo *równania różniczkowego cząstkowego* (lub PDE), które obejmuje dwie lub więcej zmiennych niezależnych.

Równania różniczkowe klasyfikuje się również ze względu na ich rząd. Na przykład równanie (8.1) nazywa się *równaniem pierwszego rzędu*, ponieważ najwyższa pochodna jest pierwszą pochodną. *Równanie drugiego rzędu* zawierałoby drugą pochodną. Na przykład równanie

$$m\frac{d^2x}{dt^2}+c\frac{dx}{dt}+kx=0, \tag{8.2}$$

opisujące położenie x układu masa-sprężyna z tłumieniem jest równaniem drugiego rzędu.

Innym przykładem klasyfikacji jest liniowość równań. Powyższe równania (8.1) i (8.2) są *równaniami liniowymi*, ze względu, że zmienna zależna i jej kolejne pochodne

w tych równaniach są wyrażone przez funkcje liniowe. Jeśli w równaniu (8.2) współczynnik oporu c zależałby od prędkości:

$$c = c_0 \frac{dx}{dt}, \tag{8.2}$$

To równanie (8.2) przyjęło by postać:

$$m\frac{d^2x}{dt^2} + c_0\left(\frac{dx}{dt}\right)^2 + kx = 0, \tag{8.3}$$

co prowadzi do *nieliniowego równania różniczkowego*.

Aby rozwiązać równania różniczkowe należy dołączyć do nich odpowiednie warunki pomocnicze. Warunki te służą do wyznaczania stałych całkowania, które powstają podczas rozwiązywania równania. Dla równania n-tego rzędu wymagane jest n warunków. Jeśli wszystkie warunki są określone przy tej samej wartości zmiennej niezależnej, wówczas mamy do czynienia z *problemem początkowym* (IVP). Z kolei istnieją problemy, w których warunki nie są znane w pojedynczym punkcie, lecz są znane dla różnych wartości zmiennej niezależnej. Ponieważ wartości te są często określane w punktach skrajnych lub na granicach układu, zazwyczaj nazywa się je *problemami brzegowymi* (BVP).

Równanie (8.1) z dołączonym warunkiem początkowym na prędkość początkową spadającego obiektu v_0 jest przykładem problemu początkowego:

$$\frac{dv}{dt} = g - \frac{c}{m}v, \; v(t=0) = v_0. \tag{8.4}$$

Z kolei równanie opisujące przepływ ciepła w pręcie o długości L z warunkami w postaci temperatur na końcach pręta jest przykładem problemu brzegowego:

$$\frac{d^2T}{dx^2} + h'(T_a - T) = 0, \; T(0) = T_0, \; T(L) = T_L, \tag{8.5}$$

gdzie h' jest współczynnikiem transferu ciepła a T_a jest temperaturą otoczenia.

8.1. Problem początkowy

Ten rozdział poświęcony jest rozwiązywaniu równań różniczkowych zwyczajnych pierwszego rzędu w formie:

$$\frac{dy}{dx} = f(x, y). \tag{8.6}$$

Metody rozwiązywania problemów początkowych opisanych równaniem (8.6) opierają się na generalnej zasadzie, którą możemy zapisać w postaci:

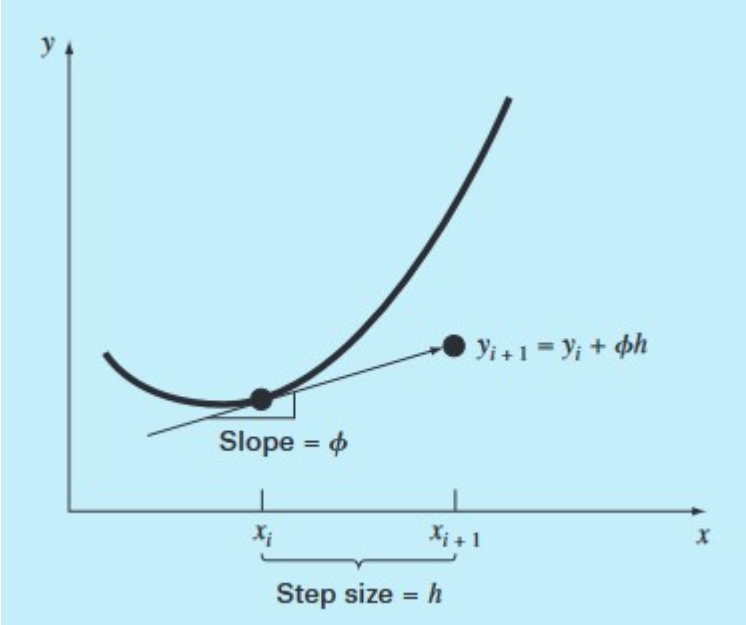
$$Nowa\ wartość = stara\ wartość + nachylenie \times krok \tag{8.7}$$

lub w zapisie matematycznym:

$$y_{i+1} = y_i + \phi h. \tag{8.8}$$

Zgodnie z równaniem (8.8), oszacowanie nachylenia funkcji ϕ służy do ekstrapolacji ze starej wartości y_i do nowej wartości y_{i+1} na odcinku h, rysunek 6. Wzór ten można stosować krok po kroku, aby obliczyć następne wartości i tym samym wyznaczyć trajektorię rozwiązania. Wszystkie metody jedno-krokowe można zapisać w tej ogólnej postaci, z jedyną różnicą jaką jest sposób szacowania nachylenia. Najprostszym podejściem jest użycie równania różniczkowego do oszacowania nachylenia w postaci pierwszej pochodnej w punkcie x_i. Innymi słowy, nachylenie na początku przedziału jest traktowane jako przybliżenie średniego nachylenia w całym przedziale h. To podejście, nazwane jest *metodą Eulera.* Metody wykorzystujące alternatywne szacowania nachylenia, prowadzące do dokładniejszych przewidywań ogólnie nazywane są *metodami Runge-Kutty.*

[Tekst alternatywny. Przedruk rysunku graficznej interpretacji metod jedno-krokowych z pozycji: Steven C. Chapra,Raymond P. Canale: „Numerical Methods for Engineers”, rys. 25.1, str. 709].



Rys. 6. Graficzna interpretacja metod jedno-krokowych.

Metoda Eulera

W metodzie Eulera nachylenie ϕ wyznacza pierwsza pochodna liczona w punkcie x_i:

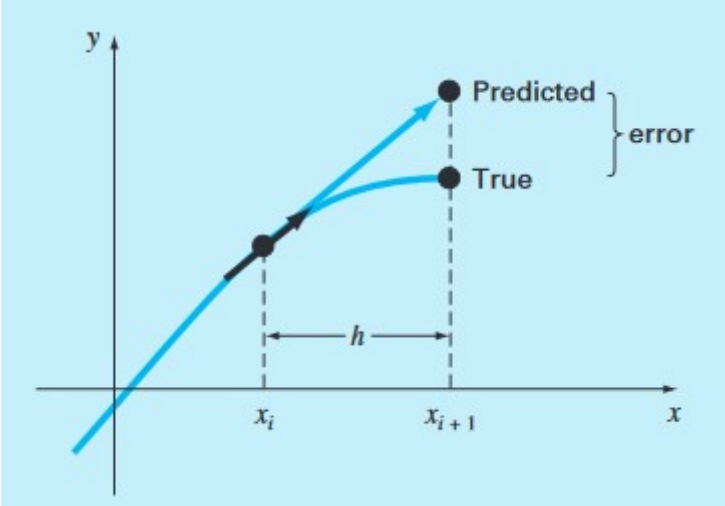
$$\phi = f(x_i, y_i), \tag{8.9}$$

gdzie $f(x_i, y_i)$ jest prawą stroną równania różniczkowego (8.6) wyznaczoną w punkcie x_i, y_i. To oszacowanie następnie jest wstawiane do równania (8.8) co prowadzi do formuły Eulera (lub inaczej formuły Eulera-Cauchy'ego):

$$y_{i+1} = y_i + f(x_i, y_i)h. \tag{8.10}$$

Nową wartość y_{i+1} przewiduje się, wykorzystując nachylenie równe pierwszej pochodnej w początkowym punkcie przedziału x_i i ekstrapoluje się je liniowo w kroku h. Graficzną interpretację metody Eulera oraz błąd metody przedstawia rysunek 7.

[Tekst alternatywny. Przedruk rysunku graficznej interpretacji metody Eulera z pozycji: Steven C. Chapra,Raymond P. Canale: „Numerical Methods for Engineers", rys. 25.2, str. 710].



Rys. 7. Graficzna interpretacja metody Eulera.

Zmniejszenie kroku h prowadzi do zmniejszenia błędu rozwiązania kosztem złożoności czasowej algorytmu, jak ma to miejsce w każdej metodzie numerycznej.

Innym sposobem zmniejszenia błędu metody Eulera może być uwzględnienie w rozwiązaniu wyrazów wyższego rzędu rozwinięcia w szereg Taylora. Na przykład, uwzględniając wyraz drugiego rzędu otrzymujemy:

$$y_{i+1} = y_i + f(x_i, y_i)h + \frac{f'(x_i, y_i)}{2!}h^2. \tag{8.11}$$

Chociaż włączenie wyrazów wyższego rzędu jest dość proste do zastosowania w przypadku kiedy prawa strona jest wielomianem, to gdy równanie różniczkowe jest

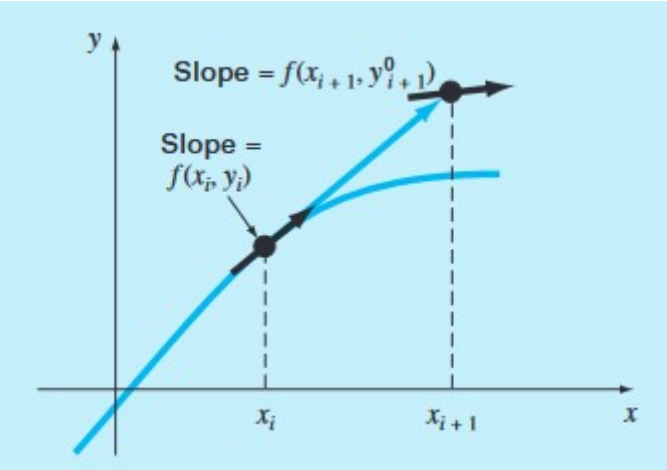
bardziej skomplikowane, w szczególności kiedy prawa strona zależy zarówno od zmiennej zależnej, jak i niezależnej, wówczas wymagane jest różniczkowanie według reguły łańcuchowej. Z tego powodu opracowano alternatywne metody jedno-krokowe. Ich schematy są porównywalne pod względem wydajności z metodami wyższego rzędu opartymi na szeregu Taylora, ale wymagają jedynie obliczenia pierwszych pochodnych.

Podstawowym źródłem błędu w metodzie Eulera jest założenie, że pochodna na początku przedziału jest taka sama w całym przedziale. Dostępna jest prosta modyfikacja nazywana *metodą Heuna*, aby poprawić dokładność metody Eulera. Jak zostanie wykazane w punkcie metoda Rungego-Kutty, modyfikacja ta w rzeczywistości należy do szerszej klasy technik rozwiązywania zwanych metodami Runge-Kutty. Ponieważ jednak ma one bardzo prostą interpretację graficzną, przedstawiona zostanie przed formalnym wyprowadzeniem jako metoda Runge-Kutty.

Metoda Heuna

Jedną z metod poprawy oszacowania nachylenia jest wyznaczenie dwóch pochodnych dla przedziału – jednej w punkcie początkowym i drugiej w punkcie końcowym. Następnie te dwie pochodne są uśredniane w celu uzyskania lepszego oszacowania nachylenia dla całego przedziału. To podejście, zwane metodą Heuna, jest przedstawione graficznie na rysunku 8.

[Tekst alternatywny. Przedruk rysunku graficznej interpretacji metody Heuna z pozycji: Steven C. Chapra,Raymond P. Canale: „Numerical Methods for Engineers”, rys. 25.9, str. 722].



Rys. 8. Graficzna interpretacja metody Heuna.

Przypomnijmy, że w metodzie Eulera nachylenie na początku przedziału:

$$y'_i = f(x_i, y_i) \tag{8.12}$$

służy do ekstrapolacji liniowej do y_{i+1}:

$$y^0_{i+1} = y_i + f(x_i, y_i)h. \tag{8.13}$$

W przypadku standardowej metody Eulera jest to ostateczne oszacowanie nachylenia w tym miejscu. Jednak w metodzie Heuna y^0_{i+1} obliczone w równaniu oszacowanie (8.13) nie jest końcowe lecz jest tylko prognozą pośrednią.

Dlatego oznaczone je indeksem górnym 0. Równanie (8.13) nazywane jest *równaniem predykcyjnym*. Dostarcza ono oszacowania y^1_{i+1}, które pozwala na obliczenie oszacowanego nachylenia na końcu przedziału.

$$y'_{i+1} = f(x_{i+1}, y^0_{i+1}). \tag{8.14}$$

W ten sposób dwa nachylenia (równania (8.12) i (8.14)) można połączyć, aby uzyskać średnie nachylenie dla przedziału:

$$\bar{y}'_{i+1} = \frac{y'_i + y'_{i+1}}{2} = \frac{f(x_i, y_i) + f(x_{i+1}, y^0_{i+1})}{2}. \tag{8.15}$$

To średnie nachylenie jest następnie wykorzystywane do ekstrapolacji liniowej z y_i do y_{i+1} przy użyciu metody Eulera:

$$y_{i+1} = y_i + \frac{f(x_i, y_i) + f(x_{i+1}, y^0_{i+1})}{2} h, \tag{8.16}$$

które nazywa się *równaniem korekcyjnym*.

Metoda Heuna jest podejściem polegającym na jedno-krokowym przewidywaniu i korygowaniu. Należy zauważyć, że ponieważ równanie (8.16) ma y_{i+1} po obu stronach znaku równości, można je stosować w sposób iteracyjny. Oznacza to, że stare oszacowanie można wielokrotnie wykorzystać do uzyskania ulepszonego oszacowania y_{i+1}. Podobnie jak w przypadku podobnych metod iteracyjnych omówionych w poprzednich punktach opracowania, kryterium zakończenia iteracji jest podane przez błąd:

$$|\varepsilon_a| = \left| \frac{y^{j+1}_{i+1} - y^j_{i+1}}{y^{j+1}_{i+1}} \right| 100\,\%, \tag{8.17}$$

gdzie y^{j+1}_{i+1} i y^j_{i+1} są rozwiązaniami z kolejnej i poprzedniej iteracji.

Metoda Rungego-Kutty

Metody Runge-Kutty (RK) osiągają dokładność porównywalną do metod wyższego rzędu opartych o szereg Taylora bez konieczności obliczania wyższych pochodnych. Istnieje wiele wariantów, ale wszystkie można przedstawić w uogólnionej postaci:

$$y_{i+1}=y_i+\phi(x_i,y_i,h)h, \tag{8.19}$$

gdzie $\phi(x_i,y_i,h)$ nazywa się funkcją inkrementacyjną, którą można interpretować jako reprezentację nachylenia afektywnego w przedziale. Funkcję inkrementacyjną można zapisać w postaci ogólnej jako:

$$\phi(x_i,y_i,h)=a_1k_1(x_i,y_i)+a_2k_2(x_i,y_i,h)+\ldots+a_nk_n(x_i,y_i,h), \tag{8.20}$$

gdzie współczynniki a_i są wartościami stałymi natomiast funkcje k_i przyjmują postać:

$$\begin{aligned}
k_1&=f(x_i,y_i)\\
k_2&=f(x_i+p_1h,y_i+q_{11}hk_1)\\
k_3&=f(x_i+p_2h,y_i+q_{12}hk_1+q_{22}hk_2)\\
&\vdots\\
k_n&=f(x_i+p_{n-1}h,y_i+q_{1n-1}hk_1+q_{2n-1}hk_2+q_{3n-1}hk_3+\ldots+q_{n-1n-1}hk_{n-1})
\end{aligned}, \tag{8.21}$$

gdzie p i q są stałymi wartościami. Należy zauważyć, że k_1 pojawia się w równaniu dla k_2, które z kolei pojawia się w równaniu dla k_3 itd., oraz że *metoda Rungego-Kuty pierwszego rzędu* dla $n=1$ prowadzi do równania metody Eulera.

Gdy $n=2$ mówimy o *metodzie Rungego-Kuty rzędu drugiego* wyrażonej równaniem (8.19) w postaci:

$$y_{i+1}=y_i+(a_1k_1+a_2k_2)h, \tag{8.22}$$

gdzie

$$\begin{aligned}
k_1&=f(x_i,y_i)\\
k_2&=f(x_i+p_1h,y_i+q_{11}hk_1)
\end{aligned}. \tag{8.23}$$

W równaniach współczynniki a_1, a_2, p_1, q_{11} wyznaczane są przez przyrównanie prawej strony wyrażenia (8.22) do szeregu Taylora drugiego rzędu:

$$y_{i+1}=y_i+f(x_i,y_i)h+\frac{f'(x_i,y_i)}{2}h^2 \tag{8.24}$$

gdzie $f'(x_i)$ wyznaczane jest z wykorzystaniem reguły łańcuchowej:

$$f'(x,y)=\frac{\partial f(x,y)}{\partial x}+\frac{\partial f(x,y)}{\partial y}\frac{dy}{dx}. \tag{8.25}$$

Podstawową strategią leżącą u podstaw metod Runge-Kutty jest wykorzystanie manipulacji algebraicznych w celu znalezienia wartości a_1, a_2, p_1 i q_{11}, które sprawią, że równania (8.24) i (8.25) będą równoważne.

Wstawiając do równania (8.24) równanie (8.25) otrzymujemy:

$$y_{i+1}=y_i+f(x_i,y_i)h+\left(\frac{\partial f}{\partial x}+\frac{\partial f}{\partial y}\frac{dy}{dx}\right)\frac{h^2}{2}. \tag{8.24}$$

Aby można było to zrobić należy rozwinąć w szereg Taylora funkcję k_2 według reguły dla dwóch zmiennych:

$$g(x+r,y+s)=g(x,y)+r\frac{\partial g}{\partial x}+s\frac{\partial g}{\partial y}+\ldots. \tag{8.25}$$

Wykorzystując wzór (8.25) do rozwinięcia funkcji k_2 wyrażonej przez drugie równanie w (8.23) otrzymujemy:

$$f(x_i+p_1h,y_i+q_{11}hk_1)=f(x_i,y_i)+p_1h\frac{\partial f}{\partial x}+q_{11}hk_1\frac{\partial f}{\partial y}. \tag{8.26}$$

Następnie równanie (8.26) razem z pierwszym równaniem z (8.23) należy wstawić do równania (8.22), co prowadzi do:

$$y_{i+1}=y_i+\left(a_1f(x_i,y_i)+a_2\left(f(x_i,y_i)+p_1h\frac{\partial f}{\partial x}+q_{11}hk_1\frac{\partial f}{\partial y}\right)\right)h. \tag{8.27}$$

Po przekształceniu otrzymujemy równanie:

$$y_{i+1}=y_i+h\left(a_1f(x_i,y_i)+a_2f(x_i,y_i)\right)+h^2\left(a_2p_1\frac{\partial f}{\partial x}+a_2q_{11}k_1\frac{\partial f}{\partial y}\right). \tag{8.28}$$

Porównując ze sobą równania (8.24) i (8.28) poprzez porównanie podobnych członów otrzymujemy trzy równania:

$$\begin{aligned}(a_1+a_2)h&=h\\ a_2p_1h^2&=\frac{h^2}{2}\\ h^2a_2q_{11}k_1&=\frac{h^2}{2}\frac{dy}{dx}\end{aligned}. \tag{8.29}$$

Zauważyć należy, że $k_1=f=\frac{dy}{dx}$, stąd aby (8.24) i (8.28) były równoważne wartości współczynników należy dobrać tak aby:

$$\begin{aligned} (a_1+a_2)&=1 \\ a_2 p_1&=\frac{1}{2} \\ a_2 q_{11}&=\frac{1}{2} \end{aligned} \quad . \tag{8.30}$$

Ponieważ w układzie (8.30) jest o jedną niewiadomą więcej niż liczba równań, nie ma unikalnego zestawu stałych, które spełniają równania. Jednak przyjmując wartość jednej ze stałych, możemy określić pozostałe trzy. W konsekwencji istnieje rodzina metod drugiego rzędu, a nie jedna wersja.

Załóżmy, że wybierzemy wartość dla zmiennej a_2, wówczas pozostałe trzy niewiadome możemy wyznaczyć bezpośrednio z układu trzech równań:

$$\begin{aligned} a_1&=1-a_2 \\ p_1&=\frac{1}{2a_2} \\ q_{11}&=\frac{1}{2a_2} \end{aligned} \quad . \tag{8.31}$$

Ponieważ możemy wybrać nieskończoną liczbę wartości dla a_2, istnieje nieskończona liczba metod Rungego-Kutty drugiego rzędu. Każda wersja dałaby dokładnie takie same wyniki, gdyby rozwiązanie równania różniczkowego zwyczajnego było kwadratowe, liniowe lub stałe. Dają one jednak różne wyniki, gdy (jak to zwykle bywa) rozwiązanie jest bardziej skomplikowane. Trzy najczęściej używane i preferowane wersje to:

- $a_2=\frac{1}{2}$ - metoda Heuna,
- $a_2=\frac{2}{3}$ - metoda Ralstona,
- $a_2=1$ - metoda punktu środkowego.

Jak można udowodnić metoda Heuna, omawiana wcześniej, jest odmianą metody Rungego-Kutty drugiego rzędu.

Wzory iteracyjne wymienionych metod można łatwo otrzymać podstawiając wartości współczynników a_1, a_2, p_1 i q_{11} do równań (8.22) i (8.23).

Postępując podobnie można wyznaczyć wzory iteracyjne dla metod wyższych rzędów. Najbardziej znaną i najczęściej używaną w praktyce jest *metoda Rungego-Kutty czwartego rzędu*. Podobnie jak w przypadku metody drugiego rzędu, istnieje nieskończenie wiele wersji metody czwartego rzędu. Poniżej znajduje się najczęściej

stosowana forma, i dlatego nazywamy ją *klasyczną metodą Rungego-Kutty czwartego rzędu*:

$$y_{i+1}=y_i+(k_1+2k_2+2k_3+k_4)h, \tag{8.32}$$

gdzie:

$$\begin{aligned} k_1&=f(x_i,y_i)\\ k_2&=f\left(x_i+\frac{1}{2}h,y_i+\frac{1}{2}hk_1\right)\\ k_3&=f\left(x_i+\frac{1}{2}h,y_i+\frac{1}{2}hk_2\right)\\ k_4&=f(x_i+h,y_i+hk_3) \end{aligned}. \tag{8.33}$$

Uwaga

Równania różniczkowe zwyczajne wyższych rzędów mogą być rozwiązywane metodami omówionymi w tym podpunkcie po zamianie na układ równań pierwszego rzędu poprzez wprowadzenie nowej zmiennej, np. równanie $m\frac{d^2x}{dt^2}+c\frac{dx}{dt}+kx=0$ możemy zapisać w postaci dwóch równań pierwszego rzędu wprowadzając nową zmienną $y=\frac{dx}{dt}$, której pochodna pierwszego rzędu jest równa $y=\frac{d^2x}{dt^2}$, stąd otrzymujemy układ równań

$$\begin{cases} \frac{dx}{dt}=y\\ \frac{dy}{dt}=\frac{-cy-kx}{m} \end{cases}.$$

Ponieważ inne równania różniczkowe n-tego rzędu można w podobny sposób zredukować, ta część opracowania koncentruje się na rozwiązywaniu równań pierwszego rzędu.

Zadanie 8.1. Rozwiąż równanie różniczkowe zwyczajne pierwszego rzędu z pochodną funkcji równą:

$$f(x,y)=-2x^3+12x^2-20x+8.5.$$

Użyj do tego celu metod:

- Eulera,
- Heuna,
- Ralstona,

- metodę punktu środkowego,
- klasyczną metodę Rungego-Kutty IV rzędu

z krokiem $h=0.5$ w przedziale $[0,4]$, z warunkiem początkowym $y(x=0)=1$.

Zaimplementuj metody w języku programowania Python, przedstaw wyniki na wykresach. Porównaj rozwiązania z rozwiązaniem dokładnym, wyznaczając błędy rzeczywiste.

Zadanie 8.2. Poniższe równanie różniczkowe zwyczajne opisuje ruch tłumionego układu sprężyna-masa:

$$m\frac{d^2x}{dt^2}+2\frac{dx}{dt}+5x=0$$

gdzie $m=1$, rozwiąż to równanie z warunkami początkowymi $\frac{dx}{dt}=0.5$ i $x=1$ w przedziale $t=[0,8]$ klasyczną metodą Rungego-Kutty IV rzędu z krokiem $h=0.5$ i $h=0.1$.

8.2. Problem brzegowy

Jeśli warunki pomocnicze do rozwiązania równania różniczkowego nie są znane w pojedynczym punkcie, lecz dla różnych wartości zmiennej niezależnej mówimy o *problemie brzegowym*.

Istnieje wiele metod aproksymacji problemu brzegowego. W niniejszym materiale omówione zostaną dwa z nich, najczęściej stosowane. Najpierw zostanie przedstawione rozwiązanie dwupunktowego problemu brzegowego *metodą różnic skończonych*. Następnie *metoda elementów skończonych* zostanie użyta do aproksymacji rozwiązania podobnego problemu.

Metoda różnic skończonych

W metodzie różnic skończonych pochodne ciągłe w równaniu różniczkowym zastępowane są wzorami różnic skończonych wyprowadzonymi w rozdziale 7. W ten sposób liniowe równanie różniczkowe przekształca się w układ równań algebraicznych, które można rozwiązać za pomocą metod omówionych w rozdziale 3.

Rozważmy problem brzegowy opisany przez zwyczajne równanie różniczkowe drugiego rzędu w formie:

$$\frac{d^2y}{dx^2}=f\left(x,y,\frac{dy}{dx}\right) \tag{8.34}$$

z warunkami brzegowymi:

$$y(a)=y_a \text{ i } y(b)=y_b, \tag{8.35}$$

gdzie a i b stanowią granice obszaru rozwiązania. Warunki brzegowe (8.35) zadane na wartość zmiennej zależnej nazywane są *warunkami podstawowymi* lub *warunkami Dirichleta*. Dalej załóżmy że problem (8.34) jest liniowy tzn., że funkcja f przyjąć może postać:

$$f(x,y,\frac{dy}{dx})=p(x)\frac{dy}{dx}+q(x)y+r(x). \tag{8.36}$$

Metoda różnic skończonych polega na zastąpieniu każdej pochodnej w równaniu różniczkowym odpowiednim przybliżeniem skończenie różnicowym. To podejście numeryczne opiera się na punktach siatki, dla których budowane są równania różnic skończonych. Siatka może być na ogół nieregularna, ale dla uproszczenia rozważymy siatkę regularną. W przypadku problemu (8.34) współrzędne węzłów są zdefiniowane przez:

$$x_i=a+ih, \text{ dla } i=0,1,2\ldots n+1, \tag{8.36}$$

gdzie $h=\frac{(b-a)}{n+1}$ jest odległością węzłów siatki skończenie różnicowej a n jest liczbą wewnętrznych węzłów. W rozważanym problemie do aproksymacji pierwszej i drugiej pochodnej zostaną użyte wzory na różnicę centralną (patrz tabela 7.1). Wzór na różnicę centralną dla pierwszej pochodnej ma postać: $\frac{y(x_{i-1})-y(x_{i+1})}{2h}$, natomiast druga pochodna wyrażona jest przez: $\frac{y(x_{i-1})-2y(x_i)+y(x_{i+1})}{h^2}$. Podstawiając te wzory do równania (8.34) otrzymujemy n równań algebraicznych w formie:

$$\frac{y(x_{i-1})-2y(x_i)+y(x_{i+1})}{h^2}=f(x_i,y(x_i),\frac{y(x_{i-1})-y(x_{i+1})}{2h}) \text{ dla } i=1,2\ldots n. \tag{8.37}$$

Aby rozwiązać ten układ należy uwzględnić warunki brzegowe (8.35) w postaci $y(x_0)=y(a)=y_0=y_a \, i \, y(x_{n+1})=y(b)=y_{n+1}=y_b$. Niewiadomymi w układzie równań (8.37) są wartości funkcji y_i w węzłach siatki skończenie różnicowej x_i.

Przykład 8.1.

Rozważmy rozwiązanie następującego problemu brzegowego metodą różnic skończonych:

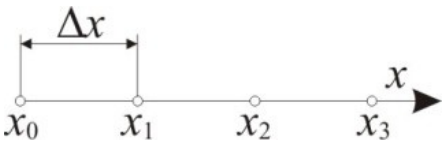
$$\frac{d^2y}{dx^2}=4(y-x) \text{ dla } 0\le x\le 1 \text{ z } y(0)=0 \, y(1)=2$$

dla którego rozwiązanie dokładne ma postać:

$$y(x)=\frac{e^2}{e^4-1}\left(e^{2x}-e^{-2x}\right)+x.$$

Do rozwiązania problemu zostanie użyta regularna siatka z odległością między węzłami $\Delta x=\frac{1}{3}$, rysunek 9.

[Tekst alternatywny. Rysunek przedstawiający siatkę skończenie różnicową dla przykładu 8.1].



Rys. 9. Siatka skończenie różnicowa dla przykładu 8.1.

Przekształcone równanie do postaci różnic skończonych dla dwóch punktów wewnętrznych, $x_1=\frac{1}{3}$ i $x_2=\frac{2}{3}$, wykorzystując wzór różnicy centralnej, przyjmuje postać dwóch równań algebraicznych:

$$\frac{y(x_2)-2y(x_1)+y(x_0)}{\Delta x^2}=4\left(y(x_1)-x_1\right)$$
$$\frac{y(x_3)-2y(x_2)+y(x_1)}{\Delta x^2}=4\left(y(x_2)-x_2\right)$$

Wstawienie warunków brzegowych $y(x_0)=0$ i $y(x_3)=2$ prowadzi do układu równań z nieznanymi wartościami $y(x_1)$ i $y(x_2)$:

$$\begin{cases}\frac{y(x_2)-2y(x_1)}{(1/3)^2}=4\left(y(x_1)-1/3\right)\\ \frac{1-2y(x_2)+y(x_1)}{(1/3)^2}=4\left(y(x_2)-2/3\right)\end{cases}$$

Rozwiązując powyższy układ równań otrzymujemy następujące wyniki: $y(x_1)=0.534$ i $y(x_2)=1.157$. Rozwiązanie dokładne w tych punktach wynosi: dla x_1 jest równe 0.531 i 1.153 dla punktu x_2. Zmniejszając odległość między węzłami Δx możemy otrzymać coraz dokładniejsze wyniki, ale kosztem wzrostu liczby równań podlegających rozwiązaniu.

Metoda elementów skończonych

Metoda elementów skończonych jest najbardziej wszechstronną i szeroko stosowaną techniką aproksymacji rozwiązań wszelkiego rodzaju problemów brzegowych. Metoda ta została pierwotnie opracowana do zastosowań w inżynierii lądowej, ale obecnie jest z powodzeniem stosowana we wszystkich dziedzinach inżynierii i matematyki. Kluczową zaletą tej techniki jest łatwość obsługi warunków brzegowych. Metoda elementów skończonych uwzględnia warunki brzegowe jako całki w sformułowaniu globalnym albo minimalizowanym funkcjonale co uniezależnia funkcje aproksymujące od specyficznych dla problemu warunków brzegowych.

Metoda elementów skończonych oparta jest na tzw. *sformułowaniu wariacyjnym* albo *globalnym*. Równanie różniczkowe stanowi tzw. *sformułowanie lokalne*, które należy przedstawić w równoważnej postaci *wariacyjnej* stosując np. *metodę Rayleigha-Ritza* lub *Bubnowa-Galerkina*. Metody te leżą u podstaw metody elementów skończonych. W niniejszych materiałach do sprowadzenia sformułowania lokalnego do równoważnego sformułowania wariacyjnego będzie użyta metoda Bubnowa-Galerkina w tzw. *sformułowaniu słabym*.

Kolejny podpunkt dotyczyć będzie rozwiązania ogólnego problemu brzegowego z użyciem metody Bubnowa-Galerkina.

Metoda Bubnowa-Galerkina

W przypadku metody Bubnowa-Galerkina, kiedy problem jest sformułowany lokalnie, rozwiązania będziemy poszukiwać budując *ważone równanie całkowe*. Takie postępowanie jest właściwe wszystkim metodom wariacyjnym rozwiązań przybliżonych. Mają one szereg zalet w porównaniu z bezpośrednim rozwiązywaniem równań różniczkowych problemu brzegowego.

Problem brzegowy wyrażony równaniem różniczkowym (czyli lokalnie) możemy przedstawić w tzw. *postaci operatorowej*:

$$Au = f \quad \text{w } \Omega, \tag{8.38}$$

gdzie A jest *operatorem różniczkowym*, u jest poszukiwaną funkcją f jest zadaną funkcją i Ω jest obszarem rozwiązania.

Załóżmy, dla uproszczenia problemu, że warunki brzegowe są jednorodne (zerowe) i podstawowe (zadane na wartość zmiennej zależnej w tym przypadku u):

$$u = 0 \quad \text{na brzegu } \Gamma. \tag{8.39}$$

Dla przykładu rozważmy równanie różniczkowe w formie:

$$-\frac{d}{dx}\left(a\frac{d}{dx}u\right)=f. \tag{8.40}$$

Operator różniczkowy równania (8.40) przyjmuje wówczas postać:

$$A=-\frac{d}{dx}\left(a\frac{d}{dx}\right). \tag{8.41}$$

Przyjmijmy pewien zbiór funkcji bazowych Ψ takich, że

$$\Psi_i \in D_A \quad i=1,2,\ldots N, \tag{8.42}$$

gdzie D_A jest przestrzenią dopuszczalnych funkcji tzn. takich, które spełniają określone warunki w obszarze Ω o których będzie mowa w dalszej części omawiania metody Bubnowa-Galerkina. Jeśli $u \in D_A$ jest taką funkcją dla której zachodzi równość:

$$(Au-f,\Psi_i)=0 \text{ w } \Omega \text{ dla każdego } i=1,2\ldots N \tag{8.43}$$

to oznacza, że u jest rozwiązaniem problemu (8.38). Operator ( ) w równaniu (8.43) oznacza iloczyn skalarny.

Innymi słowy znalezienie rozwiązania problemu (8.38) jest równoważne poszukiwaniu rozwiązania równania wariacyjnego (8.43).

Kolejnym etapem rozwiązania problemu (8.38) jest przyjęcie aproksymacji u funkcją u_N określoną na zbiorze funkcji bazowych ϕ_i z przestrzeni funkcji D_A takiej, że:

$$(Au_N-f,\Psi_i)=0 \text{ w } \Omega \text{ dla każdego } i=1,2\ldots N, \tag{8.44}$$

gdzie Au_N-f nazywane jest *residuum* (inaczej błędem rozwiązania), które przy dostatecznie dużym N dąży do dowolnie małej wartości.

Równanie (8.44) wyraża ogólną postać tzw. metod *residuów ważonych*, które zależą od postaci przyjętych baz ϕ_i i Ψ_i. Jedną z odmian metod residuów ważonych jest metoda Bubnowa-Galerkina w której ϕ_i i Ψ_i są takie same, co oznacza, że równanie (8.44) w metodzie Bubnowa-Galerkina przyjmuje postać:

$$(Au_N-f,\phi_i)=0 \text{ w } \Omega \text{ dla każdego } i=1,2\ldots N. \tag{8.45}$$

Rozwiązanie przybliżone zbudowane w oparciu o zestaw funkcji bazowych ϕ_i możemy zapisać w postaci:

$$u_N = \sum_{k=1}^{N} \phi_k c_k, \tag{8.46}$$

lub z wykorzystaniem zapisu wektorowego:

$$u_N = \boldsymbol{\phi}\, \boldsymbol{c} = \boldsymbol{c}^T \boldsymbol{\phi}^T. \tag{8.47}$$

Współczynniki c_i są wartościami podlegającymi wyznaczeniu.

Podstawiając przyjętą aproksymację funkcji u do równania (8.45) otrzymujemy:

$$\left(A \sum_{k=1}^{N} \phi_k c_k - f, \phi_i\right) = 0 \text{ w } \Omega \text{ dla każdego } i = 1,2 \ldots N. \tag{8.48}$$

Po przekształceniach równanie (4.48) możemy zapisać w postaci:

$$\left(\sum_{k=1}^{N} A\phi_k, \phi_i\right) c_k = (f, \phi_i) \text{ w } \Omega \text{ dla każdego } i = 1,2 \ldots N, \tag{8.49}$$

lub korzystając z zapisu wektorowego:

$$\left(\boldsymbol{\phi}^T, A\boldsymbol{\phi}\right) \boldsymbol{c} = \left(f, \boldsymbol{\phi}^T\right). \tag{8.50}$$

Jeżeli operator A jest operatorem dodatnio określonym, to iloczyn skalarny $\left(\boldsymbol{\phi}^T, A\boldsymbol{\phi}\right)$ możemy zapisać w postaci formy biliniowej, i wówczas równanie (8.48) przyjmuje postać:

$$\sum_{k=1}^{N} B(\phi_k, \phi_i) c_k = (f, \phi_i) + \sum_{j}^{m-n} (h_j, B_j \phi_i) \text{ w } \Omega \text{ dla każdego } i = 1,2 \ldots N, \tag{8.51}$$

gdzie składnik $\sum_{j}^{m-n} (h_j, B_j \phi_i)$ wynika z możliwości wystąpienia niejednorodnych naturalnych warunków brzegowych, w ogólnym przypadku problemu brzegowego z mieszanymi warunkami brzegowymi. W powyższym równaniu B_j jest operatorem brzegowym, związanym z podstawowym warunkiem brzegowym, h_j jest skalarem wynikającym z występującego naturalnego warunku brzegowego, $2m$ jest rzędem równania różniczkowego, a n jest liczbą wy-specyfikowanych podstawowych warunków brzegowych. W przypadku problemu z warunkami brzegowymi w postaci (8.39) człon brzegowy jest równy zero.

Sprowadzenie równania (8.49) do postaci (5.51), wyrażonej przez formę bi-liniową $B(\Phi_k, \Phi_i)$, oznacza, że rozwiązanie może być poszukiwane w przestrzeni energii H_A, czyli przestrzeni znacznie większej niż przestrzeń funkcji dopuszczalnych D_A (tzn.: wystarczy, że funkcje bazowe z tej przestrzeni są różniczkowalne do rzędu m dla

operatora różniczkowego A rzędu $2m$ i spełniają tylko jednorodne podstawowe warunki brzegowe).

Przyjmując dodatkowe oznaczenie $B(\phi_k,\phi_i)=(T\phi_k,T\phi_i)$, gdzie T jest operatorem różniczkowym rzędu m, układ równań (8.51) możemy przedstawić w postaci:

$$\sum_{k=1}^{N}(T\phi_k,T\phi_i)c_k=(f,\phi_i)\ \text{ w }\ \Omega\ \text{ dla każdego }\ i=1,2\ldots N, \tag{8.52}$$

lub w formie wektorowej:

$$\left((T\boldsymbol{\phi})^T,T\boldsymbol{\phi}\right)\boldsymbol{c}=\left(f,\boldsymbol{\phi}^{\boldsymbol{T}}\right). \tag{8.53}$$

Z postaci równania (8.53) wynika, że macierz układu równań $\boldsymbol{K}=\left((T\boldsymbol{\phi})^T,T\boldsymbol{\phi}\right)$ jest symetryczna oraz że równanie ma już postać dalej nieredukowalną. Dlatego metoda Bubnowa-Galerkina, wyrażona przez równanie (8.52), jest nazywa *słabym sformułowaniem wariacyjnym Bubnowa-Galerkina*.

Przykład 8.2.

Rozważmy problem brzegowy opisany równaniem różniczkowym:

$$\frac{d^2u}{dx^2}=-u-x\ \text{ dla }\ x\in(0,1)$$

z jednorodnymi podstawowymi warunkami brzegowymi: $u(0)=0$ i $u(1)=0$.

Rozwiązanie dokładne dla tego problemu wynosi $u_D=\dfrac{\sin(x)}{\sin(1)}-x$.

Operator różniczkowy równania A i funkcja f przyjmują postać:

$$A=\frac{d^2}{dx^2}+1,\ f=-x.$$

Zapiszmy nasze równanie w postaci wariacyjnej z wykorzystaniem metody residuów ważonych. Mnożąc residuum przez funkcję wagową w i całkując w obszarze rozwiązania otrzymujemy:

$$(Au-f,w)=\int_0^1 w\left(\frac{d^2u}{dx^2}+u+x\right)dx=0.$$

Aby problem brzegowy sprowadzić do sformułowania słabego w formie bi-liniowej $B(\Phi_k,\Phi_i)$, należy obniżyć rząd równania przez wykonanie operacji całkowania przez części, przypomnijmy ogólny wzór całkowania przez części:

$$\int_a^b w\,u''\,dx = -\int_a^b w'\,u'\,dx + w\,u'\,|_a^b.$$

Zastosowanie wzoru całkowania przez części na równaniu residuów ważonych prowadzi do równania w formie:

$$\int_0^1 -\frac{d\,w}{dx}\frac{d\,u}{dx} + wu\,dx = -\int_0^1 w\,x\,dx - w\,u'\,|_0^1,$$

gdzie $w\,u'\,|_a^b$ jest członem brzegowym zależnym od warunków brzegowych. Przypomnijmy w metodzie Bubnowa-Galerkina funkcja wagowa z założenia spełnia podstawowe warunki brzegowe, w przypadku naszego zadania $w(0)=0$ i $w(1)=0$ stąd cały człon brzegowy jest równy zero, co daje równanie w postaci:

$$\int_0^1 -\frac{d\,w}{dx}\frac{d\,u}{dx} + wu\,dx = -\int_0^1 w\,x\,dx.$$

W metodzie Bubnowa-Galerkina w sformułowaniu słabym należy przyjąć zestaw funkcji bazowych z przestrzeni H_A, tzn. takich, które są różniczkowalne do rzędu m dla operatora różniczkowego A rzędu $2m$, spełniają jednorodne podstawowe warunki brzegowe, oraz są liniowo niezależne.

Przyjmijmy dla przykładu zestaw dwóch funkcji bazowych ($N=2$). W przypadku równania różniczkowego drugiego rzędu funkcje muszą być różniczkowalne do rzędu 1, a więc mogą być np. wielomianem pierwszego rzędu.

Dla naszego zadania funkcja liniowa nie może być wybrana ponieważ spełniając warunki brzegowe ($\phi=0$) nie jest różniczkowalna do rzędu 1, dlatego możemy dobrać funkcje w postaci wielomianów wyższych rzędów np.:

$$\phi_1 = x(1-x) \text{ i } \phi_2 = x^2(1-x).$$

Zapiszmy funkcje bazowe w postaci wektora funkcji bazowych takiego, że:

$$\boldsymbol{\phi} = \begin{bmatrix} x(1-x) & x^2(1-x) \end{bmatrix}$$

i wprowadźmy aproksymację funkcji wagowej i funkcji poszukiwanej tym samym zestawem funkcji bazowych $u = \boldsymbol{\phi}\,\boldsymbol{c} = \boldsymbol{c}^T\boldsymbol{\phi}^T$ i $w = \boldsymbol{\phi}\,\boldsymbol{d} = \boldsymbol{d}^T\boldsymbol{\phi}^T$ gdzie wektory $\boldsymbol{c}$ i $\boldsymbol{d}$ mają postać:

$$\boldsymbol{c} = \begin{bmatrix} c_1 \\ c_2 \end{bmatrix},\ \boldsymbol{d} = \begin{bmatrix} d_1 \\ d_2 \end{bmatrix},$$

co prowadzi do równania w formie:

$$\int_0^1 -\boldsymbol{d}^T \frac{d\boldsymbol{\phi}^T}{dx}\frac{d\boldsymbol{\phi}}{dx}\boldsymbol{c}+\boldsymbol{d}^T\boldsymbol{\phi}^T\boldsymbol{\phi}\,\boldsymbol{c}\,dx=-\int_0^1 \boldsymbol{d}^T\boldsymbol{\phi}^T x\,dx.$$

Powyższe równanie jest spełnione dla dowolnego zestawu współczynników funkcji wagowych $\boldsymbol{d}^T\neq\boldsymbol{0}$ co prowadzi do:

$$\int_0^1 -\frac{d\boldsymbol{\phi}^T}{dx}\frac{d\boldsymbol{\phi}}{dx}+\boldsymbol{\phi}^T\boldsymbol{\phi}\,dx\,\boldsymbol{c}=-\int_0^1 \boldsymbol{\phi}^T x\,dx.$$

Podstawiając do powyższego równania wektory funkcji bazowych $\boldsymbol{\phi}$ oraz ich pochodne $\frac{d\boldsymbol{\phi}}{dx}$ w postaci:

$$\frac{d\boldsymbol{\phi}}{dx}=\begin{bmatrix}1-2x & 2x-3x^2\end{bmatrix}$$

otrzymujemy:

$$\int_0^1 \begin{bmatrix}-1+2x\\ -2x+3x^2\end{bmatrix}\begin{bmatrix}-1+2x & -2x+3x^2\end{bmatrix}+\begin{bmatrix}x(1-x)\\ x^2(1-x)\end{bmatrix}\begin{bmatrix}x(1-x) & x^2(1-x)\end{bmatrix}dx\,\boldsymbol{c}=-\int_0^1\begin{bmatrix}x(1-x)\\ x^2(1-x)\end{bmatrix}x\,dx$$

Po wymnożeniu i scałkowaniu otrzymujemy układ równań w postaci:

$$\begin{bmatrix}-0.3 & -0.15\\ -0.15 & -0.124\end{bmatrix}\begin{bmatrix}c_1\\ c_2\end{bmatrix}=\begin{bmatrix}-0.083\\ -0.05\end{bmatrix},$$

Stąd $c_1=0.192$, $c_2=0.171$. Rozwiązaniem jest funkcja w postaci:

$$u(x)=\boldsymbol{\phi}\,\boldsymbol{c}=\begin{bmatrix}x(1-x) & x^2(1-x)\end{bmatrix}\begin{bmatrix}0.192\\ 0.171\end{bmatrix}=0.192\,x(1-x)+0.171\,x^2(1-x).$$

Porównajmy uzyskane rozwiązanie z rozwiązaniem dokładnym w punkcie środkowym przedziału $x=0.5$. Wartość dokładna wynosi $u_D(0.5)=0.0697$ natomiast wartość wyznaczona metodą Bubnowa-Galerkina $u(0.5)=0.0694$. Błąd względny wynosi w tym przypadku $\varepsilon=0.434\,\%$.

Uwaga

W przypadku problemu z niejednorodnymi warunkami brzegowymi, należy *zawsze* transformować problem do problemu z jednorodnymi podstawowymi warunkami brzegowymi. Przypomnieć tu należy, że w metodzie Bubnowa-Galerkina funkcje bazowe muszą spełniać tylko podstawowe warunki brzegowe. Jeśli problem brzegowy zawiera niejednorodne *naturalne warunki brzegowe*, czyli zadane na pochodne funkcji zależnej, to takie warunki wprowadza się do równania w postaci członu brzegowego.

Rozważmy problem brzegowy z niejednorodnymi warunkami brzegowymi (podstawowym i naturalnym):

$$-y''+\pi^2 y=2\pi^2\sin(\pi x)$$

$$y(0)=1,\ y'(1)=0.5.$$

Przyjmijmy rozwiązanie w formie:

$$y=u+u_0$$

gdzie u_0 jest funkcją, którą dobieramy tak aby spełniała niejednorodny podstawowy warunek brzegowy, czyli w rozpatrywanym problemie może to być funkcja:

$$u_0=x+1$$

ponieważ $u_0(0)=1$. Funkcja u z założenie spełnia jednorodny podstawowy warunek brzegowy $u(0)=0$. Podstawiając za y przyjętą formę rozwiązania oraz za $y''=u''+u_0''$ otrzymujemy równanie:

$$-u''+\pi^2(u+x+1)=2\pi^2\sin(\pi x),$$

$$u(0)=0,\ y'(1)=u'(1)+u_0'(1)=u'(1)+1=0.5 \text{ stąd } u'(1)=-0.5,$$

reprezentujące transformowany problem do problemu z jednorodnymi podstawowymi warunkami brzegowymi. Należy pamiętać, że po rozwiązaniu tego problemu rozwiązanie problemu wyjściowego będzie uzyskane przez dodanie do funkcji u funkcji u_0.

Wprowadzenie do MES

Klasyczne metody wariacyjne (np. metoda Bubnowa-Galerkina):

- Mogą być efektywnie wykorzystane tylko do rozwiązania problemów z małą liczbą stopni swobody. Spowodowane jest to w głównej mierze z trudnościami w doborze funkcji wagowych.
- Funkcje aproksymacyjne (wagowe) zmieniają się w zależności od rozwiązywanego problemu (nie tylko zależą od równania ale także od warunków brzegowych i obszaru rozwiązania), przez to rozwiązania nie może być zautomatyzowane.

Oznacza to, że poprawa efektywności metody wariacyjnej zależy od tego czy możliwe jest, dla danej klasy problemów, konstruowanie funkcji aproksymacyjnych

dla dowolnych obszarów z wariacyjnie zgodnymi warunkami brzegowymi. Takie możliwości stwarza metoda elementów skończonych.

Metoda elementów skończonych jest procedurą wariacyjną, w której funkcje bazowe są wyznaczane w obszarze zastąpionym przez zbiór prostych pod-obszarów , na jakie został podzielony cały obszar rozwiązania. Pod-obszary te, nazywane elementami skończonymi, mają zwykle proste geometryczne kształty, co ułatwia budowanie funkcji aproksymacyjnych. Funkcje te nie zależą od danych warunków brzegowych i innych danych definiujących problem. Wszystko to powoduje, że MES w wyjątkowo dobry sposób nadaje się do obliczeń komputerowych dzięki łatwości jej zalgorytmizowania.

Rozwiązanie typowego problemu metodą elementów skończonych jest realizowane w następujących etapach:

- Podział obszaru rozwiązania na pod-obszary czego wynikiem jest zastąpienie obszaru rozwiązania zbiorem elementów skończonych (dyskretyzacja). Typ elementu zależy od obszaru i rodzaju rozwiązywanego równania różniczkowego:
  - ustalenie liczby węzłów i elementów skończonych, tworzących tzw. siatkę skończenie elementową,
  - wygenerowanie współrzędnych węzłów siatki i tablicy topologii (lub incydencji), która zawiera relację między numerami elementów i węzłów siatki.
- Wyznaczenie równań MES dla elementów:
  - sformułowanie równania wariacyjnego w obszarze elementu skończonego jedną z metod wariacyjnych (np. metodą Bubnowa-Galerkina),
  - aproksymacja nieznanych funkcji w elementach skończonych.
- Złożenie (agregacja) elementów w jeden układ, czyli budowa układu równań przy wykorzystaniu warunku zgodności zmiennych węzłowych (co oznacza, że wartości tych zmiennych we wspólnym węźle przynależnym do różnych elementów skończonych muszą być takie same). W rezultacie otrzymujemy układ równań MES całego problemu.
- Uwzględnienie warunków brzegowych, czyli wprowadzenie podstawowych i naturalnych warunków brzegowych do zagregowanego układu równań.
- Rozwiązanie układu równań ze względu na niewiadome węzłowe.

- Obliczenie dodatkowych wielkości np. obliczenie wartości funkcji rozwiązania i ich pochodnych w innych niż węzły punktach obszaru.

Z punktu widzenia matematycznego możemy interpretować metodę elementów skończonych jako rozwiązanie problemu brzegowego metodą wariacyjną (np. Bubnowa – Galerkina) z lokalnie definiowanymi funkcjami wagowymi.

Przykład rozwiązania dwu-punktowego problemu brzegowego metodą elementów skończonych

Rozważmy problem brzegowy w sformułowaniu lokalnym:

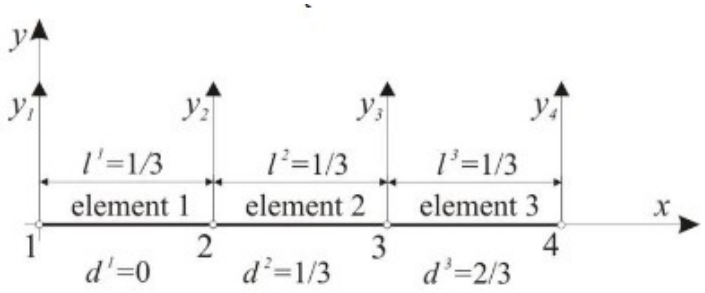
$$\frac{d^2y}{dx^2}+\frac{\pi^2}{4}y=\frac{\pi^2}{16}\cos\left(\frac{\pi}{4}\right) \text{ dla } x\in(0,1), \tag{8.54}$$

z warunkami brzegowymi:

$$y(0)=0 \quad y(1)=0. \tag{8.55}$$

Procedurę rozwiązania problemu zaczynamy od podział obszaru rozwiązania na pod-obszary czego wynikiem jest zastąpienie obszaru rozwiązania zbiorem *elementów skończonych* (dyskretyzacja), rysunek 10. Ustalamy liczbę *węzłów* i *elementów skończonych*, tworzących tzw. *siatkę skończenie elementową*.

[Tekst alternatywny. Rysunek przedstawiający siatkę skończenie elementową dla przykładu rozwiązania problemu brzegowego metodą elementów skończonych.]



Rys. 10. Dyskretyzacja obszaru.

Współrzędne węzłów przyjętej siatki skończenie elementowej zestawia tabela 8.1.

Tabela 8.1. Współrzędne węzłów siatki skończenie elementowej.

Węzeł nr	Współrzędna x
1	0
2	0.333
3	0.666
4	1

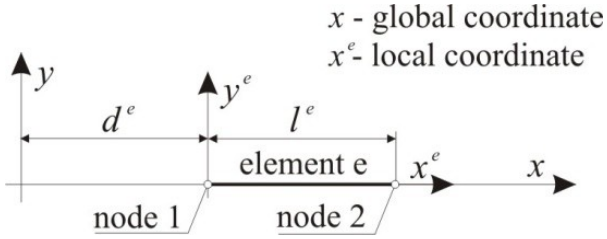
Tabela 8.2 zawiera relację między numerami elementów i węzłów siatki. Takie zestawienie nazywane jest *macierzą topologii* lub *incydencji*.

Tabela 8.2. Macierz topologii.

Element nr	Węzeł 1	Węzeł 2
1	1	2
2	2	3
3	3	4

Aby zdefiniować równanie dla elementu skończonego e, wprowadza się lokalny układ współrzędnych y^e, x^e powiązany z elementem e zdefiniowany przez parametr transformacji d^e, rysunek 11.

[Tekst alternatywny. Rysunek przedstawiający lokalny układ współrzędnych dla przykładu rozwiązania problemu brzegowego metodą elementów skończonych.]



Rys. 11. Lokalny układ współrzędnych.

Transformacja współrzędnych z układu lokalnego x^e do globalnego x jest wyrażona za pomocą relacji:

$$x = x^e + d^e. \tag{8.56}$$

Teraz równanie (8.54) można zapisać dla elementu e w jego lokalnym układzie współrzędnych, korzystając ze wzoru (8.56):

$$\frac{d^2 y^e}{d x^{e2}} + \frac{\pi^2}{4} y^e = \frac{\pi^2}{16} \cos\left(\frac{\pi}{4}\left(x^e + d^e\right)\right). \tag{8.57}$$

Metoda Bubnowa-Galerkina polega na przekształceniu problemu zapisanego w postaci lokalnej (8.57) do postaci wariacyjnej poprzez pomnożenie równania (8.57) przez funkcję testową v^e, co prowadzi do równania:

$$v^e\left(\frac{d^2 y^e}{d x^{e2}}+\frac{\pi^2}{4} y^e\right)=v^e\left(\frac{\pi^2}{16}\cos\left(\frac{\pi}{4}\left(x^e+d^e\right)\right)\right). \tag{8.58}$$

Jeżeli funkcja lokalna y^e spełnia równanie różniczkowe (8.57), całka po obszarze elementu z równania (8.58) powinna znikać:

$$\int_0^{l^e} v^e\left(\frac{d^2 y^e}{d x^{e2}}+\frac{\pi^2}{4} y^e\right)dx^e-\int_0^{l^e} v^e\left(\frac{\pi^2}{16}\cos\left(\frac{\pi}{4}\left(x^e+d^e\right)\right)\right)dx^e=0. \tag{8.59}$$

Całkowanie przez części zmniejsza rząd pochodnej o jeden co prowadzi do sformułowania słabego problemu:

$$v^e\frac{d y^e}{dx}\Bigg|_0^{l^e}+\int_0^{l^e}\frac{-d v^e}{dx}\frac{d y^e}{dx}+v^e\frac{\pi^2}{4} y^e dx^e-\int_0^{l^e} v^e\left(\frac{\pi^2}{16}\cos\left(\frac{\pi}{4}\left(x^e+d^e\right)\right)\right)dx^e=0. \tag{8.60}$$

W tym przypadku funkcje kształtu dobrano tak, aby spełniały warunek ciągłość C^1 i własność Kroneckera (5.24). Warunki te spełniają wielomiany bazowe interpolacji Lagrange'a pierwszego stopnia. Zatem wektor funkcji kształtu dla elementu ma postać:

$$\boldsymbol{N^e}\left(x^e\right)=\begin{bmatrix}N_1^e\left(x^e\right) & N_2^e\left(x^e\right)\end{bmatrix}=\begin{bmatrix}\frac{l^e-x^e}{l^e} & \frac{x^e}{l^e}\end{bmatrix}. \tag{8.61}$$

Stąd aproksymacja funkcji y^e w elemencie skończonym e wyraża się wzorem:

$$y^e\left(x^e\right)=\boldsymbol{N^e}\left(x^e\right)\boldsymbol{q^e}=\boldsymbol{q^{eT}}\boldsymbol{N^{eT}}\left(x^e\right), \tag{8.62}$$

gdzie $\boldsymbol{q^e}$ jest wektorem stopni swobody dla elementu i przyjmuje formę

$$\boldsymbol{q^e}=\begin{bmatrix}q_1^e\\ q_2^e\end{bmatrix}. \tag{8.63}$$

Do przybliżenia funkcji testowej v^e stosuje się te same funkcje kształtu, z wektorem parametrów aproksymacji $\boldsymbol{c^e}$:

$$v^e\left(x^e\right)=\boldsymbol{N^e}\left(x^e\right)\boldsymbol{c^e}=\boldsymbol{c^{eT}}\boldsymbol{N^{eT}}\left(x^e\right). \tag{8.64}$$

Pochodną funkcji y^e wyraża wzór:

$$\frac{d y^e}{d x^e}=\frac{d\boldsymbol{N^e}\left(x^e\right)}{d x^e}\boldsymbol{q^e}=\boldsymbol{q^{eT}}\frac{d\boldsymbol{N^{eT}}\left(x^e\right)}{d x^e}, \tag{8.65}$$

gdzie:

$$\frac{d\,\boldsymbol{N}^{\boldsymbol{e}}(x^e)}{d\,x^e}=\left[-\frac{1}{l^e}\quad\frac{1}{l^e}\right]. \tag{8.66}$$

Ten sam wzór stosuje się do obliczenia pochodnej funkcji testowej v^e.

Wstawiając wzory (8.62) i (8.65) do równania (8.60), pierwszy człon przekształca się do:

$$v^e\frac{d\,y^e}{dx}\bigg|_0^{l^e}=\boldsymbol{c}^T\boldsymbol{N}^{\boldsymbol{e}T}(x^e)\frac{d\,y^e}{dx}\bigg|_0^{l^e}=\boldsymbol{c}^T\begin{bmatrix}0\\1\end{bmatrix}\frac{d\,y^e}{dx}(l^e)-\boldsymbol{c}^T\begin{bmatrix}1\\0\end{bmatrix}\frac{d\,y^e}{dx}(0)=\boldsymbol{c}^T\begin{bmatrix}\dfrac{-d\,y^e}{dx}(0)\\[2ex]\dfrac{d\,y^e}{dx}(l^e)\end{bmatrix} \tag{8.67}$$

i równanie (8.60) przyjmuje postać:

$$\boldsymbol{c}^{eT}\begin{bmatrix}-\dfrac{d\,y^e}{dx^e}(0)\\[2ex]\dfrac{d\,y^e}{dx^e}(l^e)\end{bmatrix}+\boldsymbol{c}^{eT}\int_0^{l^e}-\frac{d\,\boldsymbol{N}^{eT}}{d\,x^e}\frac{d\,\boldsymbol{N}^e}{d\,x^e}+\frac{\pi^2}{4}\boldsymbol{N}^{eT}\boldsymbol{N}^e\,dx^e\,\boldsymbol{q}^e-\boldsymbol{c}^{eT}\int_0^{l^e}\boldsymbol{N}^{eT}\left(\frac{\pi^2}{16}\cos\left(\frac{\pi}{4}\left(x^e+d^e\right)\right)\right)dx^e=0 \tag{8.68}$$

Zakładając dowolne $\boldsymbol{c}^{eT}\neq\boldsymbol{0}$, ostateczna postać równania (8.60) dla elementu skończonego e w notacji macierzowej jest wyrażona wzorem:

$$\boldsymbol{K}^e\boldsymbol{q}^e=\boldsymbol{P}^e-\boldsymbol{P}^{b,e} \tag{8.69}$$

gdzie zdefiniowano odpowiednio tzw. *macierz sztywności* $\boldsymbol{K}^e$, *wektor obciążenia* $\boldsymbol{P}^e$ i *wektor graniczny* $\boldsymbol{P}^{b,e}$ w postaci:

$$\boldsymbol{K}^e=\int_0^{l^e}-\frac{d\,\boldsymbol{N}^{eT}}{d\,x^e}\frac{d\,\boldsymbol{N}^e}{d\,x^e}+\frac{\pi^2}{4}\boldsymbol{N}^{eT}\boldsymbol{N}^e\,dx^e \tag{8.70}$$

$$\boldsymbol{P}^e=\int_0^{l^e}\boldsymbol{N}^{eT}\left(\frac{\pi^2}{16}\cos\left(\frac{\pi}{4}\left(x^e+d^e\right)\right)\right)dx^e \tag{8.71}$$

$$\boldsymbol{P}^{b,e}=\begin{bmatrix}-\dfrac{d\,y^e}{dx^e}(0)\\[2ex]\dfrac{d\,y^e}{dx^e}(l^e)\end{bmatrix} \tag{8.72}$$

Dyskretyzacja przedziału rozwiązania pokazana na rysunku 10, zawiera 3 równej długości elementy skończone i 4 węzły. Ponieważ długość każdego elementu jest taka sama $l^1=l^2=l^3=\frac{1}{3}$, macierze sztywności dla elementów są również takie same.

Obliczenie całki (8.70) po długości elementu daje macierze sztywności dla elementów 1, 2 i 3 w postaci:

$$\boldsymbol{K}^1=\boldsymbol{K}^2=\boldsymbol{K}^3=-\int_0^{1/3}\begin{bmatrix}\frac{-1}{1/3}\\ \frac{1}{1/3}\end{bmatrix}\begin{bmatrix}-\frac{1}{1/3} & \frac{1}{1/3}\end{bmatrix}+\frac{\pi^2}{4}\begin{bmatrix}\frac{1/3-x^e}{1/3}\\ \frac{x^e}{1/3}\end{bmatrix}\begin{bmatrix}\frac{1/3-x^e}{1/3} & \frac{x^e}{1/3}\end{bmatrix}dx^e \tag{8.73}$$

stąd

$$\boldsymbol{K}^1=\boldsymbol{K}^2=\boldsymbol{K}^3=\begin{bmatrix}-2.726 & 3.137\\ 3.137 & -2.726\end{bmatrix}. \tag{8.74}$$

Wektory obciążenia należy obliczyć osobno, ponieważ przesunięcia układów lokalnych są różne, $d^1\neq d^2\neq d^3$. Dla elementu 1 $d_1=0$, stąd

$$\boldsymbol{P}^1=\int_0^{1/3}\begin{bmatrix}\frac{1/3-x^e}{1/3}\\ \frac{x^e}{1/3}\end{bmatrix}\left(\frac{\pi^2}{16}\cos\left(\frac{\pi}{4}x^e\right)\right)dx^e=\begin{bmatrix}0.102\\ 0.101\end{bmatrix}. \tag{8.75}$$

Dla elementu 2 $d_2=\frac{1}{3}$, stąd

$$\boldsymbol{P}^2=\int_0^{1/3}\begin{bmatrix}\frac{1/3-x^e}{1/3}\\ \frac{x^e}{1/3}\end{bmatrix}\left(\frac{\pi^2}{16}\cos\left(\frac{\pi}{4}\left(x^e+1/3\right)\right)\right)dx^e=\begin{bmatrix}0.096\\ 0.093\end{bmatrix} \tag{8.76}$$

oraz dla elementu 3 $d_3=\frac{2}{3}$

$$\boldsymbol{P}^3=\int_0^{1/3}\begin{bmatrix}\frac{1/3-x^e}{1/3}\\ \frac{x^e}{1/3}\end{bmatrix}\left(\frac{\pi^2}{16}\cos\left(\frac{\pi}{4}\left(x^e+2/3\right)\right)\right)dx^e=\begin{bmatrix}0.084\\ 0.077\end{bmatrix}. \tag{8.77}$$

Wektory brzegowe wyrażone w lokalnym układzie współrzędnych zawierają wartości pierwszych pochodnych na początku ($x^e=0$ w lokalnym układzie współrzędnych) i na końcu elementu (w $x^e=l^e$). W globalnym układzie współrzędnych wektory brzegowe przyjmują postać

$$\boldsymbol{P}^{b,1}=\begin{bmatrix}\frac{-dy}{dx}(0)\\ \frac{dy}{dx}(1/3)\end{bmatrix},\ \boldsymbol{P}^{b,2}=\begin{bmatrix}\frac{-dy}{dx}(1/3)\\ \frac{dy}{dx}(2/3)\end{bmatrix},\ \boldsymbol{P}^{b,3}=\begin{bmatrix}\frac{-dy}{dx}(2/3)\\ \frac{dy}{dx}(1)\end{bmatrix}. \tag{8.78}$$

Po obliczeniu macierzy i wektorów dla elementów, można zbudować globalny układ równań. Proces budowania równań globalnych nazywa się *agregacją*. Na podstawie numeracji węzłów i elementów wskazanych przez macierz topologii (tabela 8.2) tworzone są globalna macierz sztywności i globalny wektor brzegowy i globalny wektor obciążenia układu.

Proces agregacji macierzy sztywności rozpoczynamy od utworzenia zerowej macierzy o rozmiarze równym liczbie stopni swobody układu, w tym przypadku o rozmiarze $4\,x\,4$. Następnie do tak zbudowanej macierzy dodajemy macierze sztywności elementów skończonych w miejsca wskazane przez macierz topologii. Czyli macierz $\boldsymbol{K}^1$ zostanie dodana na miejsca (1,1), (1,2), (2,1) i (2,2), co prowadzi do:

$$\boldsymbol{K}=\begin{bmatrix} -2.726 & 3.137 & 0 & 0 \\ 3.137 & -2.726 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}. \tag{8.79}$$

Następnie dodawana jest macierz $\boldsymbol{K}^2$ w miejsca wskazane przez indeksy (2,2), (2,3), (3,2) i (3,3):

$$\boldsymbol{K}=\begin{bmatrix} -2.726 & 3.137 & 0 & 0 \\ 3.137 & -5.452 & 3.137 & 0 \\ 0 & 3.137 & -2.726 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}. \tag{8.80}$$

W ostatnim kroku budowania globalnej macierzy sztywności dodajemy macierz $\boldsymbol{K}^3$ w miejsca (3,3), (3,4), (4,3) i (4,4):

$$\boldsymbol{K}=\begin{bmatrix} -2.726 & 3.137 & 0 & 0 \\ 3.137 & -5.452 & 3.137 & 0 \\ 0 & 3.137 & -5.452 & 3.137 \\ 0 & 0 & 3.137 & -2.726 \end{bmatrix}. \tag{8.81}$$

Taką samą procedurę stosuje się do wektorów, co daje globalny wektor obciążenia i globalny wektor brzegowy w postaci:

$$\boldsymbol{P}=\begin{bmatrix} 0.102 \\ 0.197 \\ 0.177 \\ 0.079 \end{bmatrix},\ \boldsymbol{P}^b=\begin{bmatrix} -\frac{dy}{dx}(0) \\ 0 \\ 0 \\ \frac{dy}{dx}(1) \end{bmatrix}. \tag{8.82}$$

Teraz można skompletować globalny układ równań metody elementów skończonych, który stanowi model numeryczny rozwiązywanego problemu:

$$\boldsymbol{K}\,\boldsymbol{Q}=\boldsymbol{P}-\boldsymbol{P}^{b} \tag{8.83}$$

gdzie $\boldsymbol{Q}$ jest *globalnym wektorem stopni swobody* reprezentującym wartości funkcji y w węzłach siatki skończenie elementowej:

$$\boldsymbol{Q}=\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix}. \tag{8.84}$$

W postaci (8.83) układ równań jest osobliwy, aby można było go rozwiązać należy wprowadzić do niego warunki brzegowe problemu (8.55), które zadają wartości funkcji w punkcie 0 i 1 a więc znane są stopnie swobody y_1 i y_4, które przyjmują w tym przypadku zerowe wartości:

$$y(0)=y_1=0,\; y(1)=y_4=0. \tag{8.85}$$

Zatem niewiadomymi pierwotnymi w problemie są y_2 i y_3, a niewiadomymi wtórnymi są wartości pochodnych $\frac{dy}{dx}(0)$ i $\frac{dy}{dx}(1)$. Z rozwiązania tak skonstruowanego układu równań

$$\begin{bmatrix} -2.726 & 3.137 & 0 & 0 \\ 3.137 & -5.452 & 3.137 & 0 \\ 0 & 3.137 & -2.726 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}\begin{bmatrix} 0 \\ y_2 \\ y_3 \\ 0 \end{bmatrix}=\begin{bmatrix} 0.102 \\ 0.197 \\ 0.177 \\ 0.079 \end{bmatrix}-\begin{bmatrix} -\frac{dy}{dx}(0) \\ 0 \\ 0 \\ \frac{dy}{dx}(1) \end{bmatrix} \tag{8.86}$$

uzyskujemy wartości tych niewiadomych $y_2=-0.082$, $y_3=-0.08$ oraz $\frac{dy}{dx}(0)=-0.36$, $\frac{dy}{dx}(1)=0.329$.

Rozwiązanie w postaci aproksymacji funkcji w elemencie skończonym wyrażone przez globalną współrzędną x można uzyskać, korzystając z następującego wzoru:

$$y^{e}(x)=\boldsymbol{N}^{e}\left(x-d^{e}\right)\boldsymbol{q}^{e}. \tag{8.87}$$

Na przykład w elemencie 1 równanie (8.87) przyjmuje postać:

$$y^1(x)=\boldsymbol{N}^1(x-d^1)\boldsymbol{q}^1=\begin{bmatrix}\dfrac{l^1-(x-d^1)}{l^1} & \dfrac{x-d^1}{l^1}\end{bmatrix}\begin{bmatrix}q_1^1\\ q_2^1\end{bmatrix}\begin{bmatrix}0\\ -0.082\end{bmatrix}=\begin{bmatrix}\dfrac{1/3-x}{1/3} & \dfrac{x}{1/3}\end{bmatrix}\begin{bmatrix}0\\ -0.082\end{bmatrix} \tag{8.88}$$

stąd po przekształceniu otrzymujemy:

$$y^1(x)=-0.246\,x. \tag{8.89}$$

Podobnie możemy wyznaczyć funkcje aproksymujące rozwiązanie w elementach 2 i 3:

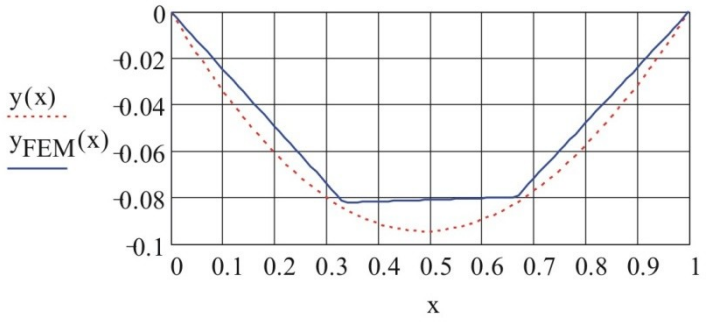
$$y^2(x)=0.006\,x-0.084, \tag{8.90}$$

$$y^3(x)=0.24\,x-0.24. \tag{8.91}$$

Wykres rozwiązania numerycznego reprezentowanego przez krzywą $y_{FEM}(x)$ i rozwiązania dokładnego reprezentowanego przez krzywą $y(x)$ przedstawiono na wykresie, rysunek 12. Ponieważ użyto aproksymacji liniowej w elementach, rozwiązanie numeryczne reprezentują odcinkowo funkcje liniowe. Rozwiązaniem dokładnym problemu (8.24) jest funkcja:

$$y(x)=\frac{-1}{3}\cos\left(\frac{\pi}{2}x\right)-\frac{\sqrt{2}}{6}\sin\left(\frac{\pi}{2}x\right)+\frac{1}{3}\cos\left(\frac{\pi}{4}x\right). \tag{8.92}$$

[Tekst alternatywny. Rysunek przedstawiający rozwiązanie numeryczne i dokładne problemu brzegowego metodą elementów skończonych.]



Rys. 12. Rozwiązanie numeryczne i dokładne problemu (8.24).

Pierwszą pochodną dla rozwiązania metodą elementów skończonych oblicza się według wzoru (8.65). Zgodnie z oczekiwaniami, pochodna funkcji liniowej daje wartości stałe:

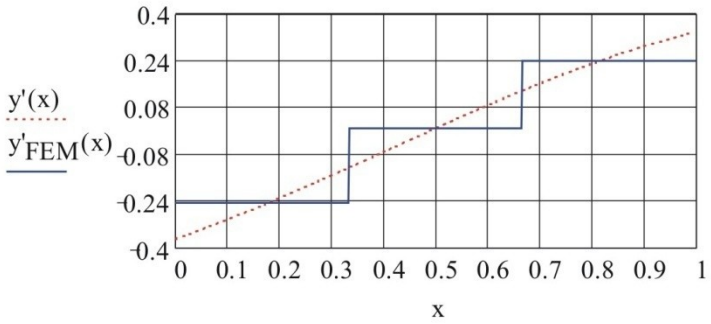
$$\frac{dy^1}{dx}(x)=\begin{bmatrix}\frac{-1}{1/3} & \frac{1}{1/3}\end{bmatrix}\begin{bmatrix}0\\-0.082\end{bmatrix}=-0.246, \tag{8.93}$$

$$\frac{dy^2}{dx}(x)=\begin{bmatrix}\frac{-1}{1/3} & \frac{1}{1/3}\end{bmatrix}\begin{bmatrix}-0.082\\-0.08\end{bmatrix}=0.006, \tag{8.94}$$

$$\frac{dy^3}{dx}(x)=\begin{bmatrix}\frac{-1}{1/3} & \frac{1}{1/3}\end{bmatrix}\begin{bmatrix}-0.08\\0\end{bmatrix}=0.24. \tag{8.95}$$

Wykres pochodnej numerycznej i dokładnej przedstawiony jest na rysunku 13.

[Tekst alternatywny. Rysunek przedstawiający pochodną rozwiązania numerycznego i dokładnego przykładowego problemu brzegowego metodą elementów skończonych.]

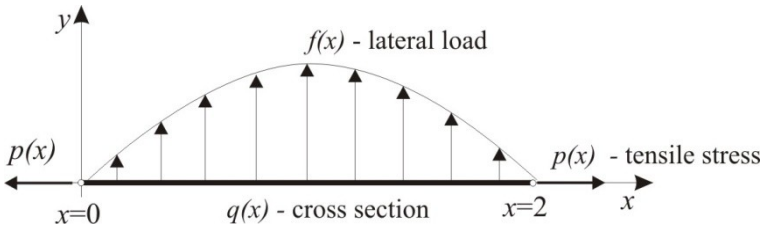


Rys. 13. Pochodna rozwiązania numerycznego i dokładnego problemu (8.24).

Jak widać, pochodna rozwiązania numerycznego jest nieciągła na granicach elementów. Nieciągłość ta wynika z przyjęcia aproksymacji liniowej. Rozwiązanie można ulepszyć, stosując aproksymację wyższego rzędu lub zwiększając liczbę elementów użytych w dyskretyzacji. Pierwsza technika nazywana jest *adaptacją typu p*, a druga *adaptacją typu h*.

Zadanie 8.3. Rozważmy model fizyczny struny o długości 2, o zmiennym przekroju poprzecznym, opisany wzorem $q(x)=0.001x+0.002$, rysunek 14. Struna jest zamocowana na obu końcach i obciążona dodatkowym naprężeniem $p(x)=100$ i rozkładem sił poprzecznych $f(x)=-x(x-2)$. Obciążenia powodują ugięcie struny. Celem jest znalezienie funkcji opisującej krzywą ugięcia struny $y(x)$.

[Tekst alternatywny. Rysunek przedstawiający model fizyczny zadania 8.3.]



Rys. 14. Model fizyczny zadania 8.3.

Silną formę modelu matematycznego przedstawia równanie różniczkowe drugiego rzędu z jednorodnymi podstawowymi warunkami brzegowymi:

$$-\frac{dy}{dx}\left(p(x)\frac{dy}{dx}\right)+q(x)\,y=f(x)\,\text{dla}\;\;0\le x\le 2,\;y(0)=0\;y(2)=0.$$

Znajdź słabą formę dla danego problemu, używając metody Bubnowa-Galerkina. Zastosuj metodę elementów skończonych do aproksymacji rozwiązania, używając 6 liniowych elementów skończonych.

Zadanie 8.4. Zasada zachowania ciepła może być wykorzystana do sporządzenia bilansu cieplnego dla długiego, cienkiego pręta. Jeżeli pręt nie jest izolowany na całej długości, a układ znajduje się w stanie ustalonym, problem definiuje równanie:

$$\frac{d^2T}{dx^2}+h'\left(T_a-T\right)=0,\;T(0)=T_0,\;T(L)=T_L,$$

gdzie h' jest współczynnikiem transferu ciepła a T_a jest temperaturą otoczenia. Znajdź rozwiązanie metodą Bubnowa-Galerkina dla dwóch funkcji bazowych dla danych: $h'=0.01$, $T_a=20$, $T_0=40$, $T_L=100$ i $L=10$. Rozwiązanie dokładne reprezentuje funkcja:

$$T=73.4523\,\mathrm{e}^{0.1x}-53.4523\,\mathrm{e}^{20.1x}+20.$$

Porównaj uzyskane wyniki do rozwiązania dokładnego.

Zadanie 8.5. Rozwiąż zadanie 8.3 używając metody elementów skończonych z dyskretyzacją obszaru rozwiązania zadaną przez liczbę elementów skończonych n. Opracuj procedury agregacji i rozwiązania układu równań wykorzystują odpowiednie biblioteki języka Python. Porównaj uzyskane wyniki do rozwiązania dokładnego.

9. Literatura

1. LeMesurier B. (2024). „Introduction to Numerical Methods and Analysis with Python”, College of Charleston, South Carolina.
2. Dudek-Dyduch E., Wąs J., Dudkiewicz L., Grobler-Dębska K., Gudowski B. (2011). „Metody numeryczne. Wybrane zagadnienia”, Wydawnictwo AGH.
3. Akai Terrence J. (1994). „Applied numerical methods for engineers”, John Wiley & Sons.
4. Chapra S., C., Canale R., P. (2015). „Numerical Methods for Engineers”, McGraw-Hill Education.

10. Spis rysunków