WSTĘP DO METOD NUMERYCZNYCH

Metodą numeryczną nazywa się każdą metodę obliczeniową sprowadzalną do operacji arytmetycznych dodawania, odejmowania, mnożenia i dzielenia. Są to podstawowe operacje matematyczne, znane od wieków przez człowieka a także rozpoznawalne przez każdy procesor komputerowy. Na fundamencie tych czterech działań liczbowych można zbudować całą bazę obliczeniową dla mniej lub bardziej skomplikowanych zagadnień (np. obliczanie pierwiastka kwadratowego z liczby nieujemnej, ale też operacje całkowania i różniczkowania numerycznego). Dlatego zazwyczaj przez *numerykę* rozumie się dziedzinę matematyki zajmującą się rozwiązywaniem przybliżonym zagadnień algebraicznych. I rzeczywiście, odkąd zjawiska przyrodnicze zaczęto opisywać przy użyciu formalizmu matematycznego, pojawiła się potrzeba rozwiązywania zadań analizy matematycznej czy algebry. Dopóki były one nieskomplikowane, dawały się rozwiązywać analitycznie, tzn. z użyciem pewnych przekształceń algebraicznych prowadzących do otrzymywania rozwiązań ścisłych danych problemów. Z czasem jednak, przy powstawaniu coraz to bardziej skomplikowanych teorii opisujących zjawiska, problemy te stawały się na tyle złożone, iż ich rozwiązywanie ścisłe było albo bardzo czasochłonne albo też zgoła niemożliwe. Numeryka pozwalała znajdywać przybliżone rozwiązania z żądaną dokładnością. Ich podstawową zaletą była ogólność tak formułowanych algorytmów, tzn. w ramach danego zagadnienia nie miało znaczenia czy było ono proste czy też bardzo skomplikowane (najwyżej wiązało się z większym nakładem pracy obliczeniowej). Natomiast wadą była czasochłonność. Stąd prawdziwy renesans metod numerycznych nastąpił wraz z powszechnym użyciem w pracy naukowej maszyn cyfrowych, a w szczególności mikrokomputerów (od lat siedemdziesiątych). Dziś złożoność metody numerycznej nie jest żadnym problemem – dziesiątki żmudnych dla człowieka operacji arytmetycznych wykonuje komputer – o wiele ważniejsza stała się analiza otrzymanego wyniku (gł. pod kątem jego dokładności) – tak, aby był on możliwie najbardziej wiarygodny.

Oczywiście metody numeryczne mogą służyć do rozwiązywania konkretnych zagadnień algebraicznych (takich jak np. równania nieliniowe czy problemy własne). Na ogół jednak są one ostatnim ogniwem w łańcuchu zwanym modelowaniem. W celu określenia zachowania się jakiegoś zjawiska w przyrodzie (tu uwaga będzie skierowana na zagadnienia fizyczne, czyli odwracalne), buduje się szereg jego przybliżeń zwanych *modelami*. Modele buduje się przyjmując coraz to nowe założenia i hipotezy upraszczające. Z rzeczywistego systemu fizycznego najpierw powstaje model mechaniczny, (czyli zbiór hipotez dotyczących np. materiału, środowiska, zachowania obciążenia itd.). Jego reprezentacją matematyczną jest model *matematyczny*, czyli opis jego zachowania się przy określonych warunkach mechanicznych w postaci układu równań różniczkowych cząstkowych (na ogół). Następny w kolejności *model numeryczny* polega na zamianie wielkości ciągłych na dyskretne – oznacza przejście do układu równań algebraicznych, do rozwiązania którego służy wybrana metoda numeryczna. Po otrzymaniu wyniku numerycznego (przybliżonego) należy przeprowadzić *analizę błędu*. Należy zauważyć, iż błąd końcowy będzie obarczony błędami ze wszystkich poprzednich etapów modelowania, a więc:

- **Błędem nieuniknionym (błędem modelu),**
- **Błędem metody,**
- **Błędem numerycznym.**

Błąd modelu zwykle wiąże się z przyjęciem złych parametrów początkowych lub brzegowych przy jego tworzeniu. Może się też okazać, iż przyjęto zbyt daleko idące uproszczenia nieoddające dobrze warunków rzeczywistych, w jakich odbywa się dane zjawisko. Mimo tego na ogół buduje się modele w miarę proste, a następnie przeprowadza *analizę wrażliwości*, tzn. sprawdza, jak duży wpływ ma dany pojedynczy czynnik na jego funkcjonowanie.
Błąd metody wiąże się z przyjęciem mało dokładnych parametrów dla tej metody (zbyt rzadki podział obszaru ciągłego na skończone odcinki) lub z zastosowaniem zbyt mało dokładnej metody (mimo dokładnych parametrów). Metod numerycznych dla danego zagadnienia jest na ogół bardzo dużo. Wybór powinien być dokonany z uwagi na przewidywaną postać rzeczywistego zachowania się zjawiska.
Błąd numeryczny wiąże się ściśle z precyzją wykonywanych obliczeń (ręcznych – przez człowieka, przez kalkulator, przez komputer). Wyróżnić można *błąd obcięcia* i *błąd zaokrągleń*. Błąd obcięcia wystąpi, gdy rozwijając daną funkcję w szereg odrzucamy nieskończoną liczbę wyrazów od pewnego miejsca, zachowując jedynie pewną początkową ich liczbę (w kalkulatorach działaniami pierwotnymi są operacje dodawania, odejmowania, mnożenia i dzielenia, natomiast wszystkie inne, np. obliczanie wartości funkcji trygonometrycznych wiąże się z rozwijaniem tychże funkcji w szeregi potęgowe z daną dokładnością obcięcia). Błąd zaokrągleń wiąże się z reprezentacją ułamków dziesiętnych nieskończonych (należy przy tym pamiętać, iż komputer prowadzi obliczenia z właściwą dla danego typu liczbowego precyzją, natomiast pokazywać graficznie wyniki może z dokładnością żądaną przez użytkownika – wtedy na potrzeby formatu prezentacji zaokrągla z daną dokładnością – tak samo jest zresztą w kalkulatorach).
Inna klasyfikacja błędu numerycznego (tu rozumianego jako dokładność) to:

- **Błąd względny (bezwymiarowy),**
- **Błąd bezwzględny.**

Przyjmując oznaczenia: $\overline{x}$ - wielkość przybliżona oraz x - wielkość ścisła, można zapisać błąd bezwzględny $\delta = |\overline{x} - x|$ i błąd względny $\varepsilon = \left|\frac{\overline{x} - x}{x}\right|$. Błąd względny jako bezwymiarowy często przedstawiany jest w procentach. Podanie samej wartości $\overline{x}$ w numeryce jest bezwartościowe – musi jej towarzyszyć jedna z powyższych dokładności, (co zapisuje się jako: $\overline{x} \pm \delta$ lub $\overline{x}(1 \pm \varepsilon)$).
Ważnym pojęciem w numeryce jest pojęcie cyfr znaczących. Pierwsza cyfra znacząca to pierwsza niezerowa cyfra licząc od lewej strony ułamka dziesiętnego. W praktyce jest to cyfra, do której można mieć „zaufanie”, iż nie pochodzi z zaokrąglenia, lecz znalazła się tam z rzeczywistych obliczeń. Np. 2345000 (4 cyfry znaczące), 2.345000 (7 cyfr znaczących), 0.023450 (5 cyfr znaczących), 0.02345 (4 cyfry znaczące) itd. Dokładność końcowa musi mieć tyle cyfr znaczących, ile mają warunki początkowe. Oznacza to w praktyce, iż nie można prowadzić obliczeń zachowując np. trzy miejsca po przecinku, a ostateczny wynik podawać bezkarnie z większą niż ta dokładnością. Będzie on wtedy bezwartościowy, gdyż błąd zaokrągleń może wkraść się nawet na pierwszą pozycję dziesiętną, zwłaszcza jeżeli w trakcie obliczeń przeprowadzano często działania dzielenia i odejmowania, które obniżają dokładność wyniku.

ROZWIĄZYWANIE NIELINIOWYCH RÓWNAŃ ALGEBRAICZNYCH

Najprostszym wykorzystaniem metod numerycznych jest numeryczne rozwiązywanie równań algebraicznych nieliniowych. Nieliniowość może być pochodzenia geometrycznego (np. w mechanice przyjęcie teorii dużych odkształceń czy przemieszczeń) lub fizycznego (nieliniowe związki konstytutywne, gdy materiał nie podlega liniowemu prawu sprężystości). Końcowym efektem takiego modelowania w przestrzeni jednowymiarowej przy jednej zmiennej niezależnej jest równanie postaci:

$$F(x) = 0$$

Tworząc w określony sposób równanie postaci:

$$x = f(x),$$

gdzie $f(x)$ jest dowolną, nieliniową funkcją zmiennej *x* można stworzyć ciąg liczbowy postaci

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

rozpoczynając obliczenia od dowolnej (na ogół) liczby x_0, zwanej *punktem startowym*:

$$x_0, \quad x_1 = f(x_0), \quad x_2 = f(x_1), \quad x_3 = f(x_2), \quad ... \tag{2}$$

Graficznie proces ten polega na szukaniu punktu wspólnego dla prostej $y = x$ oraz krzywej $y = f(x)$.
Jeżeli wykona się odpowiednio dużo takich obliczeń, to przy odpowiednich warunkach, jakie musi spełniać funkcja $f(x)$, proces okaże się zbieżny (do określonej liczby $\hat{x}$). Równanie (1) nazywa się wtedy *schematem iteracyjnym*, a ciąg przybliżeń (2) *procesem iteracyjnym*. Liczby potrzebnych iteracji nie da się z góry określić (będzie ona funkcją punktu startowego oraz postaci schematu iteracyjnego). Dlatego o miejscu przerwania iteracji muszą świadczyć dodatkowe kryteria. Formułuje się je definiując następujące nieujemne wielkości skalarne:

- Tempo zbieżności: $\varepsilon^{(1)} = \left| \dfrac{x_{n+1} - x_n}{x_{n+1}} \right|$,
- Residuum: $\varepsilon^{(2)} = \left| \dfrac{F(x_{n+1})}{F(x_0)} \right|$,
- Ilość iteracji: $\varepsilon^{(3)}: \; n = ...$

Wtedy o zakończeniu obliczeń decydować będą warunki: $\varepsilon^{(1)} \le \varepsilon^{(1)}_{dop}$, $\varepsilon^{(2)} \le \varepsilon^{(2)}_{dop}$, $n \le n_{\max}$. Dwa pierwsze są niezależne od siebie i powinny być spełnione równocześnie. Trzeci jest dla nich alternatywą. Liczby (tu: bezwymiarowe) $\varepsilon^{(1)}_{dop}$, $\varepsilon^{(2)}_{dop}$, $n_{\max}$ są danymi z góry wielkościami dopuszczalnymi.

Przy formułowaniu powyższych kryteriów użyto wielkości względnych, (które mogą być łatwo porównywane między sobą). Czasami wskazane jest użycie wielkości wymiarowych, ale wtedy określenie czy liczba jest „mała" czy „duża" nie jest już takie oczywiste.
Malejące *tempo zbieżności* świadczy o zbieżności danego schematu iteracyjnego do jednej skończonej wartości (tu: $x_n \overset{n\to\infty}{\to} \hat{x}$). Schemat iteracyjny rozbieżny może dawać coraz większe liczby wraz ze wzrostem liczby iteracji (rozbieżność jako „zbieżność" do nieskończoności), może oscylować pomiędzy dwiema różnymi wartościami (tzw. proces niestabilny) lub po prostu okazać się osobliwym dla danego x_n. Takie sytuacje wychwytuje tempo zbieżności, które zamiast systematycznie maleć utrzymuje się na tym samym poziomie lub nieograniczenie rośnie do nieskończoności.
Natomiast małość kryterium residualnego (resztkowego) świadczy o spełnieniu wyjściowego równania algebraicznego (1). Może się bowiem zdarzyć, iż sama zbieżność procesu nie gwarantuje zbieżności schematu do właściwego rozwiązania $\overline{x}$,tj. takiego, że $F(\overline{x}) = 0$. Wtedy $\hat{x} \neq \overline{x}$ i wykaże ten fakt niezerowe residuum, natomiast tempo zbieżności będzie mimo to maleć. Dopiero spełnienie obydwu kryteriów gwarantuje uzyskanie przybliżenia właściwego rozwiązania wyjściowego równania (1).

Procesy iteracyjne mogą być zbieżne i rozbieżne jednostronnie (wtedy zbliżamy się lub oddalamy od właściwego rozwiązania z jednej strony – od dołu lub od góry) lub dwustronnie (wyniki iteracji „skaczą" z jednej strony wartości ścisłej na drugą cyklicznie, przybliżając się do niej lub oddalając). Przykłady takich procesów pokazują poniższe rysunki.

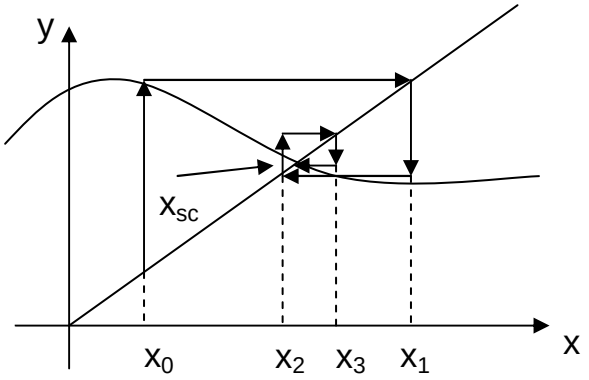
Można zauważyć pewną cechę wspólną dla funkcji prawej strony $f(x)$ w przypadku procesów zbieżnych i rozbieżnych. Wszystko zależy od nachylenia funkcji w pewnym otoczeniu (przedziale $[a,b]$), w którym poszukiwanie jest rozwiązanie. Funkcje „strome" powodują rozbieżność schematu. Tą „stromość" określa się przez kąt nachylenia wykresu do osi x, a kryterium zbieżności wynika z warunku Lipschitza.

Twierdzenie 1
Jeżeli $f(x)$ *spełnia warunek Lipschitza:*

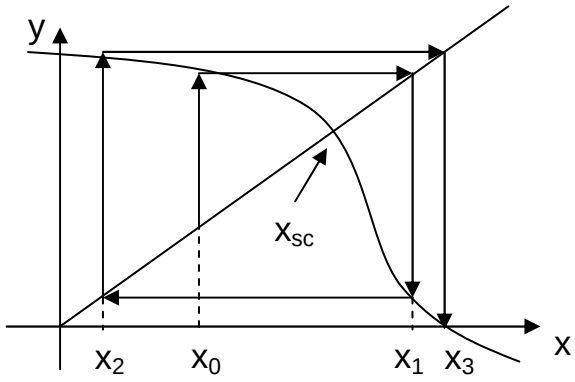
$$|f(x_1) - f(x_2)| \le L|x_1 - x_2| \quad , \quad 0 < L < 1, \quad x_1, x_2 \in [a,b]$$

to równanie algebraiczne $x = f(x)$ *posiada co najmniej jedno rozwiązanie w przedziale* $[a,b]$.

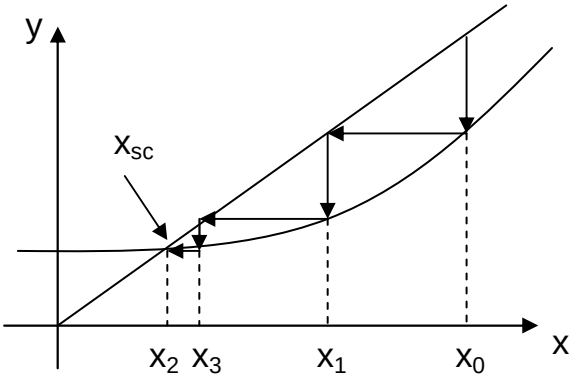
Twierdzenie 2
Jeżeli $f(x)$ *spełnia twierdzenie 1, to proces iteracyjny* $x_{n+1} = f(x_n)$ *jest zbieżny do rozwiązania ścisłego równania* $x = f(x)$ *dla* $x \in [a,b]$ *przy dowolnym punkcie startowym* x_0.



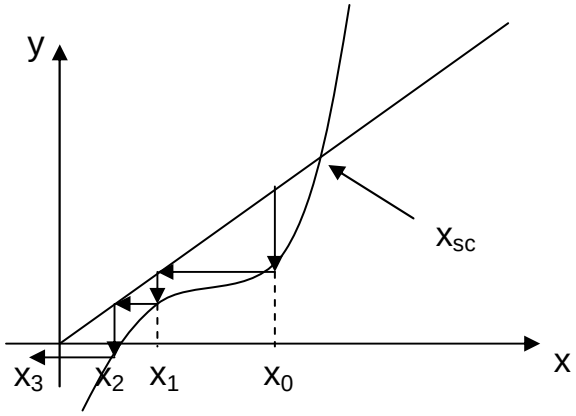
Proces zbieżny dwustronnie



Proces rozbieżny dwustronnie



Proces zbieżny jednostronnie



Proces rozbieżny jednostronnie

Konsekwencją powyższych twierdzeń jest następujący wniosek: jeżeli kąt nachylenia funkcji $f(x)$ na pewnym przedziale $x \in [x_1, x_2]$ jest mniejszy niż 45º, to schemat iteracyjny jest zbieżny przy dowolnym punkcie startowym. Tangens kąta nachylenia liczymy na podstawie ilorazu różnicowego funkcji $f(x)$.

O ewentualnej zbieżności lub rozbieżności decyduje schemat, a dokładniej: sposób znajdywania funkcji prawej strony $f(x)$. Sposób ten stanowi podstawę klasyfikacji dalszych metod iteracyjnych. Na ogół schemat powinien mieć zapewnioną bezwarunkową stabilność i zbieżność.

Równanie wyjściowe: $F(x) = 0$.

METODA ITERACJI PROSTEJ

$$\begin{cases} x_0 \\ x_{n+1} = f(x_n) \end{cases}.$$

Sposób znajdywania funkcji $f(x)$ nie jest z góry określony, może pochodzić z prostego przekształcenia: $F(x)=0 \rightarrow F(x)+x=x \rightarrow f(x)=x$. Taki schemat nie ma zagwarantowanej zbieżności ani stabilności.

METODA ITERACJI PROSTEJ Z RELAKSACJĄ

Pojęcie *relaksacji* w numeryce oznacza ingerencję w tempo zbieżności wyniku. W metodzie iteracji prostej można zrobić modyfikację polegającą na obróceniu wykresu funkcji $f(x)$ względem początku układu o taki kąt α, aby proces iteracyjny był optymalnie szybko zbieżny w okolicy danego punktu (punkt optymalnej zbieżności). Relaksacja nie tylko poprawia tempo zbieżności, ale również potrafi zamienić wyjściowy schemat rozbieżny na zbieżny.
Wartość kąta α należy wyznaczyć optymalizując nowo otrzymany schemat poprzez dodanie do starego czynnika liniowego odpowiadającego za obrót (punkt optymalnej zbieżności na ogół utożsamiany jest z punktem startowym x_0):

$$x = f(x)$$
$$x+\alpha x = f(x)+\alpha x$$
$$x(1+\alpha) = f(x)+\alpha x$$
$$x = \frac{f(x)}{1+\alpha}+\frac{\alpha x}{1+\alpha} = g(x)$$
$$x = g(x)$$

Optymalizujemy nowy schemat iteracyjny:

$$g'(x_0) = 0$$
$$\frac{f'(x_0)}{1+\alpha}+\frac{\alpha}{1+\alpha} = 0 \quad (\alpha \neq -1)$$
$$\alpha = -f'(x_0)$$

Tak policzone α wstawiamy do schematu $x = g(x)$:

$$x = \frac{f(x)}{1-f'(x_0)} - x\frac{f'(x_0)}{1-f'(x_0)}$$

Końcowa postać schematu iteracyjnego *metody iteracji prostej z relaksacją*:

$$\begin{cases} x_0 \\ x_{n+1} = \dfrac{f(x_n)}{1-f'(x_0)} - x_n\dfrac{f'(x_0)}{1-f'(x_0)} \end{cases}.$$

METODA STYCZNYCH (NEWTONA)

Dla pewnego otoczenia *h* punktu *x* rozwijamy wartość wyjściowej funkcji $F(x+h)$ w szereg Taylora:

$$F(x+h) = F(x) + F'(x)\cdot h + \frac{1}{2}F''(x)\cdot h^2 + \ldots \approx F(x) + F'(x)\cdot h$$

Ustalając x i podstawiając $F(x+h)=0$ można obliczyć przyrost h przy uprzednim odrzuceniu wyrazów rozwinięcia wyższych rzędem niż pierwszy (zlinearyzowany przyrost):

$$h = -\frac{F(x)}{F'(x)}.$$

Dla danej pary sąsiednich przybliżeń zachodzi: $x_{n+1} = x_n + h$.
Stąd po podstawieniu za *h* otrzymujemy schemat metody:

$$\begin{cases} x_0 \\ x_{n+1} = x_n - \dfrac{F(x_n)}{F'(x_n)} \end{cases}.$$

Graficznie metoda Newtona polega na budowaniu stycznych w kolejnych przybliżeniach x_n począwszy od punktu startowego oraz na szukaniu miejsc zerowych tych stycznych.
Wzór na metodę Newtona można też otrzymać stosując metodę relaksacji bezpośrednio do wyjściowego równania $F(x)=0$.
Znana jest też modyfikacja metody dla pierwiastków wielokrotnych (jeżeli równanie $F(x)=0$ takież posiada):

$$\begin{cases} x_0 \\ x_{n+1} = x_n - \dfrac{U(x_n)}{U'(x_n)}, \quad U(x) = \dfrac{F(x)}{F'(x)}, \quad U'(x) = 1 - \dfrac{F(x)\cdot F''(x)}{(F'(x))^2} \end{cases}.$$

METODA SIECZNYCH

W metodzie Newtona do schematu iteracyjnego potrzebna jest znajomość pochodnej funkcji $F(x)$. Aby uniknąć jej różniczkowania, można liczbową pochodną obliczać w sposób przybliżony korzystając z wartości ilorazu różnicowego. Wtedy potrzebne są zawsze dwa punkty wstecz, aby zbudować kolejne rozwiązanie (graficznie styczna przechodzi w sieczną), także na samym początku obliczeń.

$$\begin{cases} x_0, x_1 \\ x_{n+1} = x_n - F(x_n)\cdot \dfrac{x_n - x_{n-1}}{F(x_n) - F(x_{n-1})} \end{cases}.$$

METODA REGULA FALSI

Jeżeli zastosujemy metodę siecznych lub stycznych do funkcji nieregularnej, która w sposób gwałtowny przechodzi z wypukłej na wklęsłą lub z malejącej na rosnącą, jest

niebezpieczeństwo, iż kolejne przybliżenia rozwiązania „uciekną" daleko od początkowego przedziału bez żadnych szans na powrót i na znalezienie sensownego rozwiązania. Pomocna może się okazać pewna modyfikacja metody siecznych, gdzie jeden z punktów, na których buduje się kolejne sieczne, jest z góry ustalony (jest to jeden z punktów startowych), natomiast drugi z nich jest punktem zmiennym. W razie oddalenia się kolejnych przybliżeń od obszaru startowego, w ciągu następnych iteracji nastąpi powrót w jego okolice.

$$\begin{cases} x_0 \text{ (punkt stały)}, x_1 \\ x_{n+1} = x_n - F(x_n) \cdot \dfrac{x_n - x_0}{F(x_n) - F(x_0)} \end{cases}$$

METODA BISEKCJI (POŁOWIENIA)

Metoda bisekcji należy do najstarszych i najprostszych metod iteracyjnych, oprócz znajdywania pierwiastków równań również jest wykorzystywana przy zagadnieniach optymalizacji funkcji. Dla wyjściowego równania $F(x)=0$ szuka ona przybliżenia rozwiązania wewnątrz przedziału $x \in (a,b)$. Stąd, aby metoda zadziała, musi być gwarancja istnienia miejsca zerowego w tym przedziale: $F(a) \cdot F(b) < 0$.

Przy każdej iteracji oblicza się środek przedziału $x = \dfrac{a+b}{2}$. Następnie sprawdza się, w którym z podprzedziałów (a,x) oraz (x,b) leży miejsce zerowe i ten przedział podlega dalszemu dzieleniu. Jeżeli $F(a) \cdot F(x) < 0$ to $b = x$, w przeciwnym przypadku $a = x$. Podział przedziału (a,b) niekoniecznie musi następować na dwie równe części, można go dzielić np. w tzw. złotym stosunku (czyli tak, aby $\dfrac{b-a}{b-x} = \dfrac{b-x}{a-x}$).

Przykład 1

Podać schematy iteracyjne rozwiązania równania $\sin(x) + x^2 = 2$ metodami: (i) iteracji prostej, (ii) iteracji prostej optymalnie szybko zbieżny, (iii) Newtona, (iv) siecznych, (v) regula falsi. Zastosować tak sformułowane schematy do znalezienia dwóch kolejnych przybliżeń rozwiązania startując z punktu $x_0 = -2$ (dla metody siecznych i regula falsi przyjąć drugi punkt startowy $x_1 = -0.5$). Po każdym kroku iteracyjnym określić tempo zbieżności oraz tempo zmiany residuum.

Wyjściowe równanie: $F(x) = \sin(x) + x^2 - 2, \quad F(x) = 0$

Pierwiastki ścisłe równania: $x_{sc_1} = -1.06155, \quad x_{sc_2} = 1.728466$

(i) metoda iteracji prostej

Z równania $F(x)=0$ wyznaczamy w dowolny prosty sposób zmienną *x*, np. $x = \sqrt{\sin(x) + 2}$.

Schemat iteracyjny: $\begin{cases} x_0 = -2 \\ x_{n+1} = \sqrt{\sin(x_n) + 2}, \quad n = 0,1,2,... \end{cases}$

Obliczenia:

Krok $n = 0$:

$$x_1 = \sqrt{\sin(x_0) + 2} = \sqrt{\sin(-2) + 2} = 1.044367$$

$$e_1^{(1)} = \left| \frac{x_1 - x_0}{x_1} \right| = \left| \frac{-2 - 1.044367}{1.044367} \right| = 2.915035$$

$$e_1^{(2)} = \left| \frac{F(x_1)}{F(x_0)} \right| = \left| \frac{\sin(1.044367) - 1.044367^2 - 2}{\sin(-2) - (-2)^2 - 2} \right| = 0.609736$$

Krok $n = 1$:

$$x_2 = \sqrt{\sin(x_1) + 2} = \sqrt{\sin(1.044367) + 2} = 1.692515$$

$$e_2^{(1)} = \left| \frac{x_2 - x_1}{x_2} \right| = \left| \frac{1.044367 - 1.692515}{1.692515} \right| = 0.382950$$

$$e_2^{(2)} = \left| \frac{F(x_2)}{F(x_0)} \right| = \left| \frac{\sin(1.692515) - 1.692515^2 - 2}{\sin(-2) - (-2)^2 - 2} \right| = 0.043995$$

Z dokładnością $e_1^{(1)} = 0.000002 < 10^{-6}$ otrzymano po $n = 16$ iteracjach wynik $x_6 = 1.728466$.

(ii) metoda iteracji prostej z relaksacją

Korzystając z poprzedniego schematu metody iteracji prostej $x = f(x)$: $x = \sqrt{\sin(x) + 2}$, znajdujemy nowy schemat optymalnie szybko zbieżny w okolicy punktu startowego $x_0 = -2$.

$$f(x) = \sqrt{\sin(x) + 2} \quad \rightarrow \quad f'(x) = \frac{1}{2} \frac{1}{\sqrt{\sin(x) + 2}} \cdot \cos(x)$$

$$f'(x_0) = \frac{1}{2} \frac{1}{\sqrt{\sin(x_0) + 2}} \cdot \cos(x_0) = \frac{1}{2} \frac{1}{\sqrt{\sin(-2) + 2}} \cdot \cos(-2) = -0.199234$$

$$1 - f'(x_0) = 1.199234, \quad \frac{1}{1 - f'(x_0)} = 0.833866, \quad \frac{f'(x_0)}{1 - f'(x_0)} = -0.166134$$

Schemat iteracyjny: $\begin{cases} x_0 = -2 \\ x_{n+1} = 0.833866 \cdot \sqrt{\sin(x_n) + 2} + 0.166134 \cdot x_n, \quad n = 0,1,2,... \end{cases}$

Obliczenia:

Krok $n = 0$:

$$x_1 = 0.833866 \cdot \sqrt{\sin(x_0) + 2} + 0.166134 \cdot x_0 =$$
$$= 0.833866 \cdot \sqrt{\sin(-2) + 2} + 0.166134 \cdot (-2) = 0.538593$$
$$e_1^{(1)} = \left| \frac{x_1 - x_0}{x_1} \right| = \left| \frac{-2 - 0.538593}{0.538593} \right| = 4.713379$$
$$e_1^{(2)} = \left| \frac{F(x_1)}{F(x_0)} \right| = \left| \frac{\sin(0.538593) - 0.538593^2 - 2}{\sin(-2) - (-2)^2 - 2} \right| = 0.764049$$

Krok $n = 1$:

$$x_2 = 0.833866 \cdot \sqrt{\sin(x_1) + 2} + 0.166134 \cdot x_1 = 1.411341$$
$$e_2^{(1)} = \left| \frac{x_2 - x_1}{x_2} \right| = \left| \frac{1.411341 - 0.538593}{1.411341} \right| = 0.618382$$
$$e_2^{(2)} = \left| \frac{F(x_2)}{F(x_0)} \right| = \left| \frac{\sin(1.411341) - 1.411341^2 - 2}{\sin(-2) - (-2)^2 - 2} \right| = 0.342155$$

Z dokładnością $e_1^{(1)} = 0.000002 < 10^{-6}$ otrzymano po $n = 8$ iteracjach wynik $x_8 = 1.728464$.

(iii) metoda Newtona

Postać wyjściowa równania: $F(x) = \sin(x) + x^2 - 2, \quad F(x) = 0$.
Obliczenie pochodnej funkcji $F(x)$: $F'(x) = \cos(x) + 2x$.

Schemat iteracyjny: $\begin{cases} x_0 = -2 \\ x_{n+1} = x_n - \dfrac{\sin(x_n) + x_n^2 - 2}{\cos(x_n) + 2x_n}, \quad n = 0,1,2,... \end{cases}$

Obliczenia:

$x_1 = -1.188221, \quad e_1^{(1)} = 0.683189 \quad e_1^{(2)} = 0.116721$
$x_2 = -1.064728, \quad e_1^{(1)} = 0.115985, \quad e_1^{(2)} = 0.002854$
...
$x_4 = -1.061550, \quad e_1^{(1)} < 10^{-6}, \quad e_1^{(2)} < 10^{-8}$

(iv) metoda siecznych

Postać wyjściowa równania: $F(x) = \sin(x) + x^2 - 2, \quad F(x) = 0$.

Schemat iteracyjny:

$$\begin{cases} x_0 = -2, \quad x_1 = -0.5 \\ x_{n+1} = x_n - (\sin(x_n) + x_n^2 - 2) \cdot \dfrac{x_{n-1} - x_n}{\sin(x_{n-1}) + x_{n-1}^2 - \sin(x_n) - x_n^2}, \quad n = 1,2,... \end{cases}$$

Obliczenia:

$x_1 = -0.955962, \quad e_1^{(1)} = 0.476967 \quad e_1^{(2)} = 0.092554$
$x_2 = -1.078578, \quad e_1^{(1)} = 0.113683, \quad e_1^{(2)} = 0.015336$
...
$x_5 = -1.061550, \quad e_1^{(1)} < 10^{-6}, \quad e_1^{(2)} < 10^{-8}$

(v) metoda regula falsi

Postać wyjściowa równania: $F(x) = \sin(x) + x^2 - 2, \quad F(x) = 0$.
Punkt stały: $x_0 = -2$.

Schemat iteracyjny:

$$\begin{cases} x_0 = -2, \quad x_1 = -0.5 \\ x_{n+1} = x_n - (\sin(x_n) + x_n^2 - 2) \cdot \dfrac{-2 - x_n}{3.090703 - \sin(x_n) - x_n^2}, \quad n = 1, 2, ... \end{cases}$$

Obliczenia:
$x_1 = -0.955962, \quad e_1^{(1)} = 0.476967 \quad e_1^{(2)} = 0.092554$
$x_2 = -1.044406, \quad e_1^{(1)} = 0.084684, \quad e_1^{(2)} = 0.015327$
...
$x_7 = -1.061548, \quad e_1^{(1)} < 10^{-6}, \quad e_1^{(2)} < 10^{-6}$

Przykład 2
Równanie z poprzedniego zadania rozwiązać w sposób przybliżony metodą bisekcji. Przyjąć przedział $(1,3)$. Rozwiązanie znaleźć z dokładnością $e_{dop} = 10^{-3}$.

Postać wyjściowa równania: $F(x) = \sin(x) + x^2 - 2, \quad F(x) = 0$.
Początek przedziału: $a_0 = 1$, koniec przedziału: $b_0 = 3$.

Obliczenia zestawiono w tabeli:

n	$x_n = \frac{a_{n-1} + b_{n-1}}{2}$	$F(x_n) \cdot F(a_{n-1})$	a_n	b_n	$e_n^{(1)} = \left\|\frac{x_{n-1} - x_n}{x_n}\right\|$	$\delta_n^{(2)} = \|F(x_n)\|$
1	2.000	-2.008497	1.000	2.000	0.500	1.090703
2	1.500	1.376490	1.500	2.000	0.333333	0.747495
3	1.750	-0.058689	1.500	1.750	0.142857	0.078514
4	1.625000	0.267533	1.625000	1.750	0.076923	0.357906
5	1.687500	0.052090	1.687500	1.750	0.037037	0.145542
6	1.718750	0.005090	1.718750	1.750	0.018182	0.034973
7	1.734375	-0.000749	1.718750	1.734375	0.009009	0.021406
8	1.726563	0.000240	1.726563	1.734375	0.004525	0.006875
9	1.730469	-0.000050	1.726563	1.730469	0.002257	0.007243
10	1.728516	-0.000001	1.726563	1.728516	0.001130	0.000178
11	**1.727539**	0.000023	1.727539	1.728516	**0.000565**	**0.003350**

UKŁADY RÓWNAŃ NIELINIOWYCH

Rozwiązywanie układów równań algebraicznych (liniowych lub nieliniowych) to najczęściej spotykany problem algebraiczny w zagadnieniach fizyki. Stąd potrzeba opracowania aparatu analizy takich układów, najczęściej w formie wektorowej i macierzowej. Ponieważ działania wykonywane będą już nie na pojedynczych liczbach tylko na wielkościach macierzowych, należy wprowadzić pojęcie normy (wektora, macierzy) – stanowiącej reprezentację tej wielkości w postaci pojedynczej liczby rzeczywistej dodatniej.

Definicja
Norma wektorowa $\|\boldsymbol{x}\|$ *z wektora* $\boldsymbol{x} \in V$, *gdzie* V *to liniowa n – wymiarowa przestrzeń wektorowa, jest skalarem spełniającym następujące warunki:*

1. $\|\boldsymbol{x}\| \geq 0 \quad \forall_{\boldsymbol{x}\in V} \quad , \quad \|\boldsymbol{x}\| = 0 \quad \Leftrightarrow \quad \boldsymbol{x} = \boldsymbol{0}$,
2. $\|\alpha \boldsymbol{x}\| = |\alpha| \cdot \|\boldsymbol{x}\| \quad \forall_{\boldsymbol{x}\in V} \quad , \quad \forall_{\alpha \in \Re}$,
3. $\|\boldsymbol{x} + \boldsymbol{y}\| \leq \|\boldsymbol{x}\| + \|\boldsymbol{y}\| \quad \forall_{\boldsymbol{x},\boldsymbol{y}\in V}$.

Najczęściej używane normy wektorowe:

1. $\|\boldsymbol{x}\|_1 = \left[\sum_{i=1}^{n} |x_i|^2 \right]^{\frac{1}{2}}$, norma Euklidesa (średnio kwadratowa),
2. $\|\boldsymbol{x}\|_2 = \max_{(i)} |x_i|$, norma Czebyszewa (maksimum),
3. $\|\boldsymbol{x}\|_3 = \left[\sum_{i=1}^{n} |x_i|^p \right]^{\frac{1}{p}}$, $p \geq 1$, uogólnienie dwóch powyższych przypadków ($p = 2$ - norma Euklidesa, $p = \infty$ - norma Czebyszewa).

Definicja
Norma macierzowa $\|\boldsymbol{A}\|$ *z macierzy kwadratowej* $n \times n$ ($\underset{n\times n}{\boldsymbol{A}} = \left[a_{ij} \right]$) *jest skalarem spełniającym następujące warunki:*

1. $\|\boldsymbol{A}\| \geq 0 \quad , \quad \|\boldsymbol{A}\| = 0 \quad \Leftrightarrow \quad \boldsymbol{A} = \boldsymbol{0}$,
2. $\|\alpha \boldsymbol{A}\| = |\alpha| \cdot \|\boldsymbol{A}\| \quad , \quad \forall_{\alpha \in \Re}$,
3. $\|\boldsymbol{A} + \boldsymbol{B}\| \leq \|\boldsymbol{A}\| + \|\boldsymbol{B}\|$,
4. $\|\boldsymbol{A} \cdot \boldsymbol{B}\| \leq \|\boldsymbol{A}\| \cdot \|\boldsymbol{B}\|$.

Najczęściej używane normy macierzowe:

1. $\|\boldsymbol{A}\|_1 = \left[\sum_{i=1}^{n} \sum_{j=1}^{n} |a_{ij}|^2 \right]^{\frac{1}{2}}$, norma Euklidesa (średnio kwadratowa),
2. $\|\boldsymbol{A}\|_2 = \max_{(i)} \sum_{j=1}^{n} |a_{ij}|$, norma Czebyszewa (maksimum).

Często używane jest też pojęcie średniej normy Euklidesa. Wtedy przed pierwiastkowaniem sumy kwadratów współrzędnych dzieli się dodatkowo tą sumę przez liczbę wyrazów *n*.

METODA NEWTONA – RAPHSONA

Metoda służy do rozwiązywania układów równań nieliniowych i stanowi uogólnienie metody iteracji prostej dla wielu równań jednocześnie.

Twierdzenie 1
Niech $F_i : x_i \in [a_i, b_i] \to \Re, \quad i = 1, 2, ..., n$ *należy do n – wymiarowej przestrzeni euklidesowej* $\Re^n$.
Niech $\boldsymbol{x} = \boldsymbol{f}(\boldsymbol{x})$ *spełnia następujące warunki*

1. $\boldsymbol{f}$ *jest określone i ciągłe w* $\Re^n$,
2. *norma jakobianowa z* $\boldsymbol{f}$ *spełnia warunek* $\left\| \boldsymbol{J}_f(\boldsymbol{x}) \right\| \le L \le 1$,

$$\boldsymbol{J}_f = \begin{bmatrix} \frac{\partial F_1}{\partial x_1} & \dots & \frac{\partial F_1}{\partial x_n} \\ \dots & \dots & \dots \\ \frac{\partial F_n}{\partial x_1} & \dots & \frac{\partial F_n}{\partial x_n} \end{bmatrix}$$

3. *dla każdego* $\boldsymbol{x} \in \Re^n$ *m* $\boldsymbol{f}(\boldsymbol{x})$ *również należy do* $\Re^n$.

Wtedy dla każdego $\boldsymbol{x_0} \in \Re^n$ *ciąg iteracyjny* $\boldsymbol{x}_{n+1} = \boldsymbol{f}(\boldsymbol{x}_n)$ *jest zbieżny do jednoznacznego rozwiązania* $\tilde{\boldsymbol{x}}$.

Rozważmy punkt $\boldsymbol{x}$ oraz jego bliskie otoczenie $\boldsymbol{x} + \boldsymbol{h}$. Wtedy $\boldsymbol{F}(\boldsymbol{x} + \boldsymbol{h}) = \boldsymbol{0}$. Rozwijając ostatnią wielkość wektorową w szereg Taylora otrzymuje się:

$$\boldsymbol{F}(\boldsymbol{x} + \boldsymbol{h}) = \boldsymbol{F}(\boldsymbol{x}) + \frac{\partial \boldsymbol{F}(\boldsymbol{x})}{\partial \boldsymbol{x}} \boldsymbol{h} + \frac{1}{2} \frac{\partial^2 \boldsymbol{F}(\boldsymbol{x})}{\partial^2 \boldsymbol{x}} \boldsymbol{h}^2 + ... = \boldsymbol{F}(\boldsymbol{x}) + \boldsymbol{J}(\boldsymbol{x}) \cdot \boldsymbol{h} + \boldsymbol{R}(\boldsymbol{x}) = \boldsymbol{0}$$

Linearyzując powyższy związek ze względu na $\boldsymbol{h}$ i wyliczając wektor $\boldsymbol{h}$ otrzymuje się:

$$\boldsymbol{F}(\boldsymbol{x}) + \boldsymbol{J}(\boldsymbol{x}) \cdot \boldsymbol{h} = \boldsymbol{0} \quad \to \quad \boldsymbol{h} = -\boldsymbol{J}^{-1}(\boldsymbol{x}) \cdot \boldsymbol{F}(\boldsymbol{x})$$

$$\boldsymbol{x}_{n+1} = \boldsymbol{x}_n + \boldsymbol{h} \quad \to \quad \boldsymbol{x}_{n+1} = \boldsymbol{x}_n - \boldsymbol{J}^{-1}(\boldsymbol{x}_n) \cdot \boldsymbol{F}(\boldsymbol{x}_n)$$

Przy takim zapisie schematu konieczne byłoby odwracanie macierzy $\boldsymbol{J}^{-1}(\boldsymbol{x}_n)$ na każdym kroku. Aby tego uniknąć, mnoży się stronami przez $\boldsymbol{J}(\boldsymbol{x}_n)$, co prowadzi do sformułowania układu równań liniowych (rozwiązywanych analitycznie lub numerycznie).

$$\begin{cases} \boldsymbol{x_0} \\ \boldsymbol{J}(\boldsymbol{x}_n) \cdot \boldsymbol{x}_{n+1} = \boldsymbol{J}(\boldsymbol{x}_n) \cdot \boldsymbol{x}_n - \boldsymbol{F}(\boldsymbol{x}_n) \end{cases}$$

Kryteria przerwania iteracji w przypadku wielowymiarowym:

1. $\varepsilon^{(1)} = \dfrac{\left\| \boldsymbol{x}_{n+1} - \boldsymbol{x}_n \right\|}{\left\| \boldsymbol{x}_{n+1} \right\|} \le \varepsilon_{dop}^{(1)}$,
2. $\varepsilon^{(2)} = \dfrac{\left\| \boldsymbol{F}(\boldsymbol{x}_{n+1}) \right\|}{\left\| \boldsymbol{F}(\boldsymbol{x}_0) \right\|} \le \varepsilon_{dop}^{(2)}$.

Najczęściej sprawdza się rozwiązanie dla dwóch rodzajów norm: dla normy maksimum, która wychwytuje największy błąd w obszarze rozwiązania i średniej normy kwadratowej, która mówi o średniej jakości rozwiązania.

Istnieją różne modyfikacje metody Newtona. Najprostsza polega na nie zmienianiu wyjściowej macierzy jakobianowej, co pociąga większą liczbę kroków, niż przy oryginalnej metodzie, ale tylko dla jednego obliczania macierzy (pomocne może być omawiane w dalszych rozdziałach opracowania rozbicie na czynniki trójkątne). Możliwe jest też uaktualnianie macierzy co pewną liczbę kroków, a więc tam, gdzie macierz mogła ulec istotnej zmianie.
Inna metoda polega na dokonaniu *relaksacji*.

$$\begin{cases} \boldsymbol{x_0} \\ \boldsymbol{J}(\boldsymbol{x_n}) \cdot \boldsymbol{x_{n+1}} = \boldsymbol{J}(\boldsymbol{x_n}) \cdot \boldsymbol{x_n} - \alpha \cdot \boldsymbol{F}(\boldsymbol{x_n}) \end{cases}.$$

(najczęściej $\alpha = 1.2, 1.3, 1.4$ - tzw. *nadrelaksacja*)

W przypadku wyraźnej oscylacji rozwiązania (np. wynik przechodzi z jednej na drugą stronę osi „x") możliwe jest wprowadzenie przyśpieszenia zbieżności iteracji *metodą Aitkena*. Wprowadzając oznaczenia: x_j - wartość rozwiązania na j-tym kroku, x - rozwiązanie ścisłe można zapisać liniowy związek:

$$x - x_n = \alpha(x - x_{n-1})$$

Przy założeniu, że współczynnik α jest stały dla dwóch sąsiednich iteracji, można zapisać układ równań dla trzech kolejnych przybliżeń rozwiązania:

$$\begin{cases} x - x_n = \alpha(x - x_{n-1}) \\ x - x_{n-1} = \alpha(x - x_{n-2}) \end{cases}$$

Rozwiązując go ze względu na x otrzymuje się związek:

$$x = \frac{x_{n-2} \cdot x_n - x_{n-1}^2}{x_n - 2x_{n-1} + x_{n-2}}.$$

Wzór należy używać osobno dla każdej zmiennej niezależnej poprawiając wartość otrzymaną na n-tym kroku iteracji.

Przykład 1

Rozwiązać następujący układ równań nieliniowych $\begin{cases} y^2 = 2x \\ x^2 + y^2 = 8 \end{cases}$ metodą Newtona – Raphsona. Przyjąć wektor startowy $\boldsymbol{x}_0 = \left\{0, 2\sqrt{2}\right\}$. Po każdym kroku iteracyjnym przeprowadzać analizę błędu. Przyjąć następujące poziomy błędu: $\varepsilon_{dop}^{(1)} = 10^{-6}$, $\varepsilon_{dop}^{(2)} = 10^{-8}$.

Wektor funkcyjny: $\boldsymbol{F}(x,y) = \begin{cases} F_1(x,y) = y^2 - 2x \\ F_2(x,y) = x^2 + y^2 - 8 \end{cases}$.

Macierz jakobianowa: $\boldsymbol{J}(x,y) = \begin{bmatrix} \frac{\partial F_1}{\partial x} & \frac{\partial F_1}{\partial y} \\ \frac{\partial F_2}{\partial x} & \frac{\partial F_2}{\partial y} \end{bmatrix}(x,y) = \begin{bmatrix} -2 & 2y \\ 2x & 2y \end{bmatrix}$.

Wektor startowy: $\boldsymbol{x}_0 = \begin{bmatrix} 0 \\ 2\sqrt{2} \end{bmatrix} = \begin{bmatrix} 0.0 \\ 2.8284 \end{bmatrix}$.

Schemat iteracyjny: $\boldsymbol{J}(x_n, y_n) \cdot \boldsymbol{x}_{n+1} = \boldsymbol{J}(x_n, y_n) \cdot \boldsymbol{x}_n - \boldsymbol{F}(x_n, y_n) \quad \rightarrow \quad \boldsymbol{x}_{n+1} = \ldots$

$$\begin{bmatrix} -2 & 2y_n \\ 2x_n & 2y_n \end{bmatrix} \cdot \begin{bmatrix} x_{n+1} \\ y_{n+1} \end{bmatrix} = \begin{bmatrix} -2 & 2y_n \\ 2x_n & 2y_n \end{bmatrix} \cdot \begin{bmatrix} x_n \\ y_n \end{bmatrix} - \begin{bmatrix} y_n^2 - 2x_n \\ x_n^2 + y_n^2 - 8 \end{bmatrix} \quad \rightarrow \quad \begin{bmatrix} x_{n+1} \\ y_{n+1} \end{bmatrix} = \ldots$$

Analiza błędów:

$$\varepsilon^{(1)} = \frac{\left\|\boldsymbol{x}_{n+1} - \boldsymbol{x}_n\right\|}{\left\|\boldsymbol{x}_{n+1}\right\|} = \frac{\left\|\begin{bmatrix} x_{n+1} - x_n \\ y_{n+1} - y_n \end{bmatrix}\right\|}{\left\|\begin{bmatrix} x_{n+1} \\ y_{n+1} \end{bmatrix}\right\|} \overset{?}{<} \varepsilon_{dop}^{(1)}, \quad \varepsilon^{(2)} = \frac{\left\|\boldsymbol{F}(\boldsymbol{x}_{n+1})\right\|}{\left\|\boldsymbol{F}(\boldsymbol{x}_0)\right\|} = \frac{\left\|\begin{bmatrix} y_n^2 - 2x_n \\ x_n^2 + y_n^2 - 8 \end{bmatrix}\right\|}{\left\|\begin{bmatrix} y_0^2 - 2x_0 \\ x_0^2 + y_0^2 - 8 \end{bmatrix}\right\|} \overset{?}{<} \varepsilon_{dop}^{(2)}$$

Krok $n = 0$:

$$\begin{bmatrix} -2 & 2y_0 \\ 2x_0 & 2y_0 \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ y_1 \end{bmatrix} = \begin{bmatrix} -2 & 2y_0 \\ 2x_0 & 2y_0 \end{bmatrix} \cdot \begin{bmatrix} x_0 \\ y_0 \end{bmatrix} - \begin{bmatrix} y_0^2 - 2x_0 \\ x_0^2 + y_0^2 - 8 \end{bmatrix} \quad \rightarrow \quad \begin{bmatrix} x_1 \\ y_1 \end{bmatrix} = \ldots$$

$$\begin{bmatrix} -2.0 & 5.6569 \\ 0.0 & 5.6569 \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ y_1 \end{bmatrix} = \begin{bmatrix} -2.0 & 5.6569 \\ 0.0 & 5.6569 \end{bmatrix} \cdot \begin{bmatrix} 0.0 \\ 2.8284 \end{bmatrix} - \begin{bmatrix} 8.0 \\ 0.0 \end{bmatrix}$$

$$\begin{bmatrix} -2.0 & 5.6569 \\ 0.0 & 5.6569 \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ y_1 \end{bmatrix} = \begin{bmatrix} 8.0 \\ 16.0 \end{bmatrix} \rightarrow \quad \begin{bmatrix} \boldsymbol{x_1} \\ \boldsymbol{y_1} \end{bmatrix} = \begin{bmatrix} \boldsymbol{4.0} \\ \boldsymbol{2.8284} \end{bmatrix}$$

Błędy w normie euklidesowej:

$$e_e^{(1)} = \frac{\left\|\boldsymbol{x}_1 - \boldsymbol{x}_0\right\|_e}{\left\|\boldsymbol{x}_1\right\|_e} = \frac{\left\|\begin{bmatrix}4.0\\2.8284\end{bmatrix} - \begin{bmatrix}0.0\\2.8284\end{bmatrix}\right\|_e}{\left\|\begin{bmatrix}4.0\\2.8284\end{bmatrix}\right\|_e} = \frac{\left\|\begin{bmatrix}4.0\\0.0\end{bmatrix}\right\|_e}{\left\|\begin{bmatrix}4.0\\2.8284\end{bmatrix}\right\|_e} = \frac{\sqrt{\frac{1}{2}\left(4.0^2 + 0.0^2\right)}}{\sqrt{\frac{1}{2}\left(4.0^2 + 2.8284^2\right)}} = 0.8165$$

$$e_e^{(2)} = \frac{\left\|\boldsymbol{F}(\boldsymbol{x}_1)\right\|_e}{\left\|\boldsymbol{F}(\boldsymbol{x}_0)\right\|_e} = \frac{\left\|\begin{bmatrix}2.8284^2 - 2\cdot 4.0\\4.0^2 + 2.8284^2 - 8.0\end{bmatrix}\right\|_e}{\left\|\begin{bmatrix}2.8284^2 - 2\cdot 0.0\\0.0^2 + 2.8284^2 - 8.0\end{bmatrix}\right\|_e} = \frac{\left\|\begin{bmatrix}0.0\\16.0\end{bmatrix}\right\|_e}{\left\|\begin{bmatrix}8.0\\0.0\end{bmatrix}\right\|_e} = \frac{\sqrt{\frac{1}{2}\left(0.0^2 + 16.0^2\right)}}{\sqrt{\frac{1}{2}\left(8.0^2 + 0.0^2\right)}} = 2.0$$

Błędy w normie maksimum:

$$e_m^{(1)} = \frac{\left\|\boldsymbol{x}_1 - \boldsymbol{x}_0\right\|_m}{\left\|\boldsymbol{x}_1\right\|_m} = \frac{\left\|\begin{bmatrix}4.0\\2.8284\end{bmatrix} - \begin{bmatrix}0.0\\2.8284\end{bmatrix}\right\|_m}{\left\|\begin{bmatrix}4.0\\2.8284\end{bmatrix}\right\|_m} = \frac{\left\|\begin{bmatrix}4.0\\0.0\end{bmatrix}\right\|_m}{\left\|\begin{bmatrix}4.0\\2.8284\end{bmatrix}\right\|_m} = \frac{4.0}{4.0} = 1.0$$

$$e_m^{(2)} = \frac{\left\|\boldsymbol{F}(\boldsymbol{x}_1)\right\|_m}{\left\|\boldsymbol{F}(\boldsymbol{x}_0)\right\|_m} = \frac{\left\|\begin{bmatrix}2.8284^2 - 2\cdot 4.0\\4.0^2 + 2.8284^2 - 8.0\end{bmatrix}\right\|_m}{\left\|\begin{bmatrix}2.8284^2 - 2\cdot 0.0\\0.0^2 + 2.8284^2 - 8.0\end{bmatrix}\right\|_m} = \frac{\left\|\begin{bmatrix}0.0\\16.0\end{bmatrix}\right\|_m}{\left\|\begin{bmatrix}8.0\\0.0\end{bmatrix}\right\|_m} = \frac{16.0}{8.0} = 2.0$$

Sprawdzenie kryterium zbieżności:

$$\varepsilon_e^{(1)} = 0.8165 > \varepsilon_{dop}^{(1)} = 10^{-6}$$
$$\varepsilon_m^{(1)} = 1.0000 > \varepsilon_{dop}^{(1)} = 10^{-6}$$
$$\varepsilon_e^{(2)} = 2.0000 > \varepsilon_{dop}^{(2)} = 10^{-8}$$
$$\varepsilon_m^{(2)} = 2.0000 > \varepsilon_{dop}^{(2)} = 10^{-8}$$

Krok $n = 1$:

$$\begin{bmatrix}-2 & 2y_1\\2x_1 & 2y_1\end{bmatrix}\cdot\begin{bmatrix}x_2\\y_2\end{bmatrix} = \begin{bmatrix}-2 & 2y_1\\2x_1 & 2y_1\end{bmatrix}\cdot\begin{bmatrix}x_1\\y_1\end{bmatrix} - \begin{bmatrix}y_1^2 - 2x_1\\x_1^2 + y_1^2 - 8\end{bmatrix} \rightarrow \begin{bmatrix}x_2\\y_2\end{bmatrix} = \ldots$$

$$\begin{bmatrix}-2.0 & 5.6569\\8.0 & 5.6569\end{bmatrix}\cdot\begin{bmatrix}x_1\\y_1\end{bmatrix} = \begin{bmatrix}-2.0 & 5.6569\\8.0 & 5.6569\end{bmatrix}\cdot\begin{bmatrix}4.0\\2.8284\end{bmatrix} - \begin{bmatrix}0.0\\16.0\end{bmatrix} = \begin{bmatrix}8.0\\16.0\end{bmatrix} \rightarrow \begin{bmatrix}\boldsymbol{x_2}\\\boldsymbol{y_2}\end{bmatrix} = \begin{bmatrix}\boldsymbol{2.40}\\\boldsymbol{2.2627}\end{bmatrix}.$$

Błędy w normie euklidesowej:

$$e_e^{(1)}=\frac{\|\boldsymbol{x}_2-\boldsymbol{x}_1\|_e}{\|\boldsymbol{x}_2\|_e}=\frac{\left\|\begin{bmatrix}2.40\\2.2627\end{bmatrix}-\begin{bmatrix}4.0\\2.8284\end{bmatrix}\right\|_e}{\left\|\begin{bmatrix}2.40\\2.2627\end{bmatrix}\right\|_e}=0.5145,\quad e_e^{(2)}=\frac{\|\boldsymbol{F}(\boldsymbol{x}_2)\|_e}{\|\boldsymbol{F}(\boldsymbol{x}_0)\|_e}=\frac{\left\|\begin{bmatrix}2.2627^2-2\cdot 2.40\\2.40^2+2.2627^2-8.0\end{bmatrix}\right\|_e}{\left\|\begin{bmatrix}8.0\\0.0\end{bmatrix}\right\|_e}=0.6667$$

Błędy w normie maksimum:

$$e_m^{(1)}=\frac{\|\boldsymbol{x}_2-\boldsymbol{x}_1\|_m}{\|\boldsymbol{x}_2\|_m}=\frac{\left\|\begin{bmatrix}2.40\\2.2627\end{bmatrix}-\begin{bmatrix}4.0\\2.8284\end{bmatrix}\right\|_m}{\left\|\begin{bmatrix}2.40\\2.2627\end{bmatrix}\right\|_m}=0.3622,\quad e_m^{(2)}=\frac{\|\boldsymbol{F}(\boldsymbol{x}_2)\|_m}{\|\boldsymbol{F}(\boldsymbol{x}_0)\|_m}=\frac{\left\|\begin{bmatrix}2.2627^2-2\cdot 2.40\\2.40^2+2.2627^2-8.0\end{bmatrix}\right\|_m}{\left\|\begin{bmatrix}8.0\\0.0\end{bmatrix}\right\|_m}=0.3600$$

Sprawdzenie kryterium zbieżności:

$$\varepsilon_e^{(1)}=0.5145>\varepsilon_{dop}^{(1)}=10^{-6}$$
$$\varepsilon_m^{(1)}=0.6667>\varepsilon_{dop}^{(1)}=10^{-6}$$
$$\varepsilon_e^{(2)}=0.3622>\varepsilon_{dop}^{(2)}=10^{-8}$$
$$\varepsilon_m^{(2)}=0.3600>\varepsilon_{dop}^{(2)}=10^{-8}$$

Krok $n=2$:

$$\begin{bmatrix}-2.0 & 4.5255\\4.80 & 4.5255\end{bmatrix}\cdot\begin{bmatrix}x_3\\y_3\end{bmatrix}=\begin{bmatrix}-2.0 & 4.5255\\4.80 & 4.5255\end{bmatrix}\cdot\begin{bmatrix}2.40\\2.2627\end{bmatrix}-\begin{bmatrix}0.320\\2.880\end{bmatrix}=\begin{bmatrix}5.120\\18.88\end{bmatrix}\rightarrow\begin{bmatrix}\boldsymbol{x_3}\\\boldsymbol{y_3}\end{bmatrix}=\begin{bmatrix}\boldsymbol{2.0235}\\\boldsymbol{2.0257}\end{bmatrix}$$

Błędy w normie euklidesowej: $e_e^{(1)}=\frac{\|\boldsymbol{x}_3-\boldsymbol{x}_2\|_e}{\|\boldsymbol{x}_3\|_e}=0.1554,\quad e_e^{(2)}=\frac{\|\boldsymbol{F}(\boldsymbol{x}_3)\|_e}{\|\boldsymbol{F}(\boldsymbol{x}_0)\|_e}=0.1859$

Błędy w normie maksimum: $e_m^{(1)}=\frac{\|\boldsymbol{x}_3-\boldsymbol{x}_2\|_m}{\|\boldsymbol{x}_3\|_m}=0.0257,\quad e_m^{(2)}=\frac{\|\boldsymbol{F}(\boldsymbol{x}_3)\|_m}{\|\boldsymbol{F}(\boldsymbol{x}_0)\|_m}=0.0247$

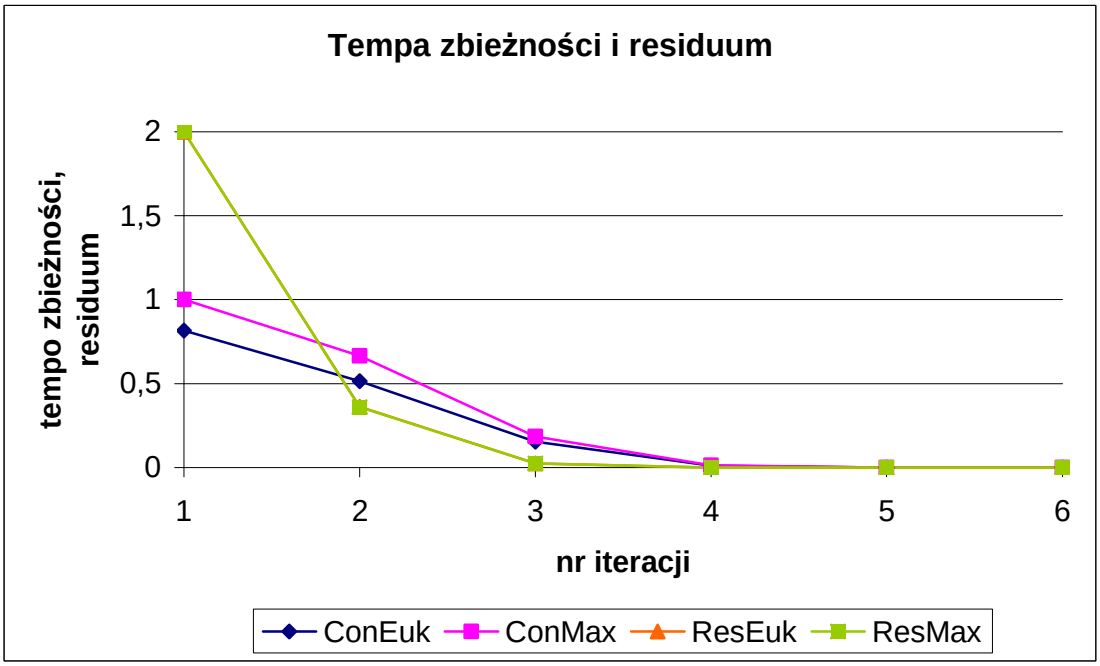
Sprawdzenie kryterium zbieżności:

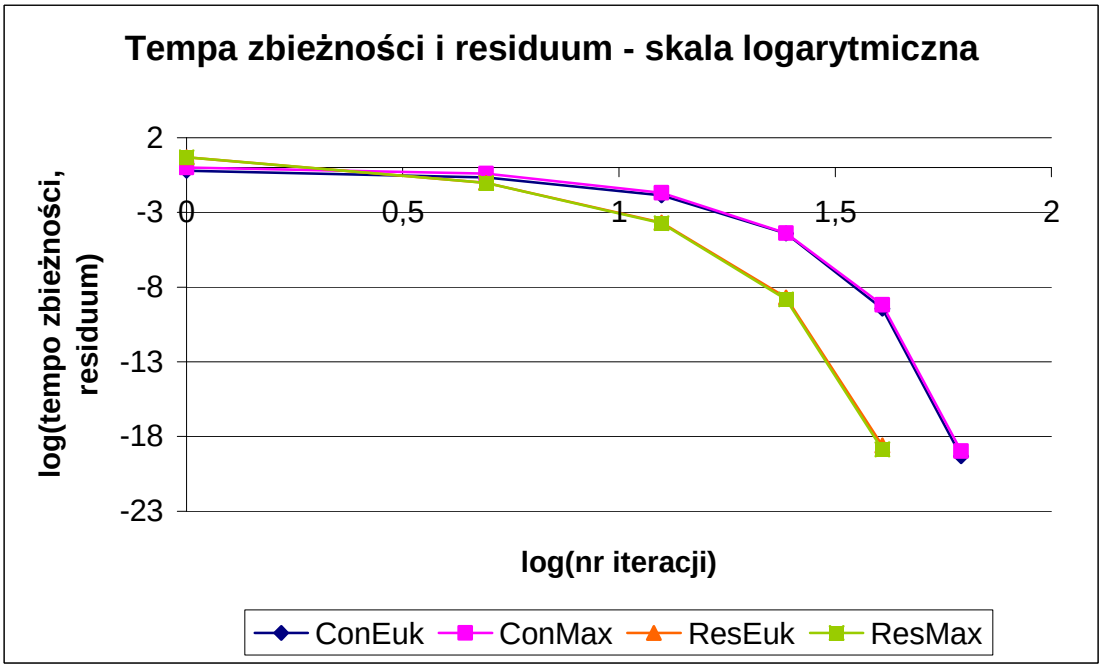
$$\varepsilon_e^{(1)}=0.1554>\varepsilon_{dop}^{(1)}=10^{-6}$$
$$\varepsilon_m^{(1)}=0.1859>\varepsilon_{dop}^{(1)}=10^{-6}$$
$$\varepsilon_e^{(2)}=0.0257>\varepsilon_{dop}^{(2)}=10^{-8}$$
$$\varepsilon_m^{(2)}=0.0247>\varepsilon_{dop}^{(2)}=10^{-8}$$

itd.

Wyniki spełniające kryteria zbieżności otrzymano po szóstej iteracji $\boldsymbol{x}_6 = \{x_6 = 2.000, y_6 = 2.0000\}$.

Poniższe wykresy przedstawiają tempo zbieżności rozwiązania i residuum: w skali dziesiętnej i logarytmicznej liczone dla obydwu powyżej zastosowanych norm.





Przyśpieszenie zbieżności metodą Aitkena ma sens wtedy, gdy rozwiązanie wyraźnie „skacze", przechodząc cyklicznie z jednej strony na drugą pewnej ustalonej wartości. W przypadku powyższym wyraźnie obserwowana jest zbieżność „od góry", a więc włączenie algorytmu Aitkena nie jest uzasadnione i może popsuć dobre już rozwiązania. Od strony formalnej jego zastosowanie będzie polegało na obliczeniu nowej, poprawionej wartości rozwiązania po trzecim kroku iteracyjnym.

Rozwiązanie uzyskane po trzecim kroku: $\begin{bmatrix} x_3 \\ y_3 \end{bmatrix} = \begin{bmatrix} 2.0235 \\ 2.0257 \end{bmatrix}$

Poprawienie współrzędnej x_3: $\overline{x}_3 = \dfrac{x_1 \cdot x_3 - x_2^2}{x_3 - 2x_2 + x_1} = \dfrac{4.0 \cdot 2.0235 - 2.40^2}{2.0235 - 2 \cdot 2.40 + 4.0} = 1.9985$

Poprawienie współrzędnej y_3: $\overline{y}_3 = \dfrac{y_1 \cdot y_3 - y_2^2}{y_3 - 2y_2 + y_1} = \dfrac{2.2627 \cdot 2.0257 - 2.8284^2}{2.0257 - 2 \cdot 2.8284 + 2.2627} = 1.9971$

Następny krok iteracyjny ($n = 3$) uwzględniałby oczywiście już poprawione powyżej wartości rozwiązania:

$$\begin{bmatrix} -2.0 & 3.9943 \\ 3.9971 & 3.9943 \end{bmatrix} \cdot \begin{bmatrix} x_4 \\ y_4 \end{bmatrix} = \begin{bmatrix} -2.0 & 3.9943 \\ 3.9971 & 3.9943 \end{bmatrix} \cdot \begin{bmatrix} 1.9985 \\ 1.9971 \end{bmatrix} + \begin{bmatrix} 0.0085 \\ 0.0173 \end{bmatrix} = \begin{bmatrix} 3.9886 \\ 15.9827 \end{bmatrix} \rightarrow \begin{bmatrix} \boldsymbol{x_4} \\ \boldsymbol{y_4} \end{bmatrix} = \begin{bmatrix} \boldsymbol{2.0000} \\ \boldsymbol{2.0000} \end{bmatrix}$$

Wartości rozwiązania po tym kroku z dokładnością do sześciu miejsc po przecinku równają się wynikowi ścisłemu.

UKŁADY RÓWNAŃ LINIOWYCH

W metodzie Newtona – Rhapsona po linearyzacji równań należy rozwiązać układ równań liniowych. Sposób algebraiczny (metoda Cramera) wymaga liczenia wyznaczników i jest dość kłopotliwy. Dlatego wprowadzono szereg metod numerycznych do rozwiązywania takich układów równań.

Rozważany będzie następujący problem algebry:

$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}, \quad \det(\boldsymbol{A}) \neq 0$

$\boldsymbol{A}$ - macierz współczynników układu ($n \times n$),
$\boldsymbol{x}$ – wektor rozwiązań ($n \times 1$),
$\boldsymbol{b}$ – wektor prawej strony (wyrazy wolne) ($n \times 1$).

Klasyfikacja metod do rozwiązywania powyższego zagadnienia może opierać się na własnościach macierzy współczynników. Wtedy można rozróżnić:

1. macierz symetryczną: $\boldsymbol{A}^T = \boldsymbol{A}$,
2. macierz dodatnio określoną: $\boldsymbol{x}^T \boldsymbol{A}\boldsymbol{x} > 0, \quad \forall_{\boldsymbol{x} \in \Re^n}$,
3. macierz o dużym rozmiarze: $n >> 1$,
4. macierz o specjalnej strukturze (np. pasmowej).

Metody rozwiązywania można podzielić wtedy na:

- metody eliminacji (polegają na odpowiednim rozkładzie macierzy **A** na czynniki a następnie na wyliczeniu jednego po drugim wszystkich rozwiązań) – są uciążliwe obliczeniowo, ale za to dają wynik ścisły, np. metoda Gaussa – Jordana, metoda Choleskiego;
- metody iteracyjne (polegają na zastosowaniu prostych metod iteracyjnych do każdego z równań algebraicznych z osobna, co daje w rezultacie ciąg wektorów przybliżeń rozwiązania ścisłego), np. metoda Jacobiego, metoda Gaussa – Seidela, metoda Richardsona;
- metody kombinowane (eliminacyjno – iteracyjne);
- metody specjalne, np. metody analizy frontalnej czy metody macierzy rzadkich (macierz ma wiele zer, mało współczynników niezerowych, np. metoda Thomasa).

Metody eliminacji, które zostaną omówione poniżej, polegają na rozkładzie wyjściowej macierzy **A** na czynniki, tzw. czynniki trójkątne **L** i **U**: $\boldsymbol{A} = \boldsymbol{L} \times \boldsymbol{U}$. Macierz dolnotrójkątna $\boldsymbol{L}$ ma następującą własność: współczynniki niezerowe występują jedynie poniżej wyrazów na przekątnej głównej, tj. $\underset{(n\times n)}{\boldsymbol{L}} : l_{ij} = \begin{cases} \neq 0, & j \le i \\ 0, & j > i \end{cases}$, macierz górnotrójkątna $\boldsymbol{U}$ ma własność

odwrotną: współczynniki niezerowe położone są powyżej przekątnej głównej, tj. $\underset{(n\times n)}{\boldsymbol{U}} : u_{ij} = \begin{cases} 0, & j < i \\ \neq 0, & j \le i \end{cases}$. Po znalezieniu tego rozkładu rozwiązuje się tzw. „pozorne” układy równań: krok wprzód: $\boldsymbol{Ly} = \boldsymbol{b}$ oraz krok wstecz: $\boldsymbol{Ux} = \boldsymbol{y}$. Układy, dzięki swojej trójkątnej strukturze, pozwalają na uzyskanie kolejnych rozwiązań rekurencyjnie wiersz po wierszu zaczynając liczenie od góry (przy macierzy dolnotrójkątnej) lub od dołu (przy macierzy górnotrójkątnej).

METODA GAUSSA – JORDANA

$$\boldsymbol{Ax} = \boldsymbol{b}, \quad \det(\boldsymbol{A}) \neq 0$$

$$\sum_{j=1}^{n} a_{ij} x_j = b_i, \quad i = 1, 2, ..., n$$

Wzory dla wersji eliminacji elementów pod przekątną główną i krokiem wstecz:
$\boldsymbol{Ab} \rightarrow \boldsymbol{Uy} \rightarrow \boldsymbol{Ux} = \boldsymbol{y} \rightarrow \boldsymbol{x}$

$$\begin{aligned} a_{ij}^{(k)} &= a_{ij}^{(k-1)} - m_{ik} \cdot a_{kj}^{(k-1)} \\ b_i^{(k)} &= b_i^{(k-1)} - m_{ik} \cdot b_k^{(k-1)} \end{aligned}$$, gdzie: $m_{ik} = \frac{a_{ik}^{(k-1)}}{a_{kk}^{(k-1)}}, \quad k = 1, 2, ..., n-1; \quad i = k+1, ..., n; \quad j = 1, ..., n$

$$x_i = \left[b_i - \sum_{j=i+1}^{n} a_{ij} x_j \right] \cdot \frac{1}{a_{ii}}, \quad i = n, n-1, ..., 2, 1$$

Wzory dla wersji eliminacji elementów nad przekątną główną i krokiem wprzód: ***Ab*** $\rightarrow$ ***Ly*** $\rightarrow$ ***Lx = y*** $\rightarrow$ ***x***

$$\begin{aligned} a_{ij}^{(k)} &= a_{ij}^{(k-1)} - m_{ik} \cdot a_{kj}^{(k-1)} \\ b_i^{(k)} &= b_i^{(k-1)} - m_{ik} \cdot b_k^{(k-1)} \end{aligned}, \text{ gdzie: } m_{ik} = \frac{a_{ik}^{(k-1)}}{a_{kk}^{(k-1)}}, \quad k = n, n-1, ..., 2; \quad i = k-1, ..., 1; \quad j = 1, ..., n$$

$$x_i = \left[b_i - \sum_{j=1}^{i-1} a_{ij} x_j \right] \cdot \frac{1}{a_{ii}}, \quad i = 1, 2, ..., n$$

Wzory dla wersji pełnej eliminacji elementów macierzy: ***Ab*** $\rightarrow$ ***Uy*** $\rightarrow$ ***Lx*** $\rightarrow$ ***x***

$$\begin{aligned} a_{ij}^{(k)} &= a_{ij}^{(k-1)} - m_{ik} \cdot a_{kj}^{(k-1)} \\ b_i^{(k)} &= b_i^{(k-1)} - m_{ik} \cdot b_k^{(k-1)} \end{aligned}, \text{ gdzie: } m_{ik} = \frac{a_{ik}^{(k-1)}}{a_{kk}^{(k-1)}}, \quad k = 1, 2, ..., n-1; \quad i = k+1, ..., n; \quad j = 1, ..., n$$

$$\begin{aligned} a_{ij}^{(k)} &= a_{ij}^{(k-1)} - m_{ik} \cdot a_{kj}^{(k-1)} \\ b_i^{(k)} &= b_i^{(k-1)} - m_{ik} \cdot b_k^{(k-1)} \end{aligned}, \text{ gdzie: } m_{ik} = \frac{a_{ik}^{(k-1)}}{a_{kk}^{(k-1)}}, \quad k = n, n-1, ..., 2; \quad i = k-1, ..., 1; \quad j = 1, ..., n$$

$$x_i = \frac{b_i}{a_{ii}}, \quad i = 1, 2, ..., n$$

W przypadku, gdy przy $\det(\boldsymbol{A}) \neq 0$, a mimo to przy obliczaniu współczynnika m_{ik} wyraz $a_{kk}^{(k-1)} = 0$ należy odwrócić kolejność wierszy (o numerach ”i” oraz „k”) tablicy złożonej z wyrazów macierzy współczynników oraz wyrazów wektora prawej strony. Można też rozwiązywać układy równań z wieloma prawymi stronami, wtedy całą macierz prawych stron ($\underset{(n \times m)}{\boldsymbol{B}} = [b_{ij}]$, gdzie *m* jest liczbą prawych stron) przetwarza się równocześnie.

Przykład 1

Rozwiązać metodą eliminacji Gaussa – Jordana układ równań $\boldsymbol{Ax} = \boldsymbol{b}$, gdzie $\boldsymbol{A} = \begin{bmatrix} 1 & -2 & -1 \\ -2 & 6 & 3 \\ -1 & 3 & 10 \end{bmatrix}$, $\boldsymbol{b} = \begin{bmatrix} -6 \\ 19 \\ 35 \end{bmatrix}$.

Zastosowana zostanie wersja pełnej eliminacji wyrazów macierzy do postaci diagonalnej. Z wyrazów macierzy ***A*** oraz wyrazów wektora ***b*** budujemy tablicę liczb:

$$\left| \begin{array}{cccc} 1 & -2 & -1 & -6 \\ -2 & 6 & 3 & 19 \\ -1 & 3 & 10 & 35 \end{array} \right|.$$

W pierwszym kroku eliminacji podlegają elementy pod przekątną główną (z macierzy ***A*** powstanie macierz górnotrójkątna ***U***), kolejno „-2”, „”-1” oraz „3”. Do zerowania „-2”

używamy czynnika eliminacji $m_{21}=\frac{-2}{1}=-2$. Jest on równy ilorazowi wyrazu, który ma się wyzerować („-2") przez odpowiadający mu wyraz stojący w pierwszym wierszu od góry, który nie ulega zmianie (tu: wiersz pierwszy, wyraz „1"). Następnie zmianie podlega każdy wyraz w wierszu drugim (łącznie z ostatnią kolumną wyrazów wolnych) wg przepisu: nowy wyraz równa się różnicy starego wyrazu i iloczynu współczynnika „m" przez wyraz z tej samej kolumny z wiersza górnego niezmiennego dla tego kroku (znowu wiersz pierwszy).
Stąd nowa postać wiersza drugiego:
$a_{21}=-2-(-2)\cdot 1=-2+2=0$, $a_{22}=6-(-2)\cdot(-2)=6-4=2$, $a_{23}=3-(-2)\cdot(-1)=3-2=1$,
$b_2=19-(-2)\cdot(-6)=19-12=7$.

Podobnie dla wyzerowania wyrazu $a_{31}=-1$ współczynnik $m_{31}=\frac{-1}{1}=-1$ a nowy zestaw wyrazów:

$a_{31}=-1-(-1)\cdot 1=-1+1=0$, $a_{32}=3-(-1)\cdot(-2)=3-2=1$, $a_{33}=10-(-1)\cdot(-1)=10-1=9$,
$b_3=35-(-1)\cdot(-6)=35-6=29$.

Tablica wyrazów po tym kroku wygląda następująco:

$$\begin{vmatrix} 1 & -2 & -1 & -6 \\ 0 & 2 & 1 & 7 \\ 0 & 1 & 9 & 29 \end{vmatrix}.$$

Cały proces sprowadza się tak naprawdę do pomnożenia pierwszego równania najpierw przez „-2" i dodaniu go do drugiego a następnie przez „-1" i dodaniu go do trzeciego.

W następnym, ostatnim „górnotrójkątnym" kroku eliminacji podlega „1" (dawne „3"). Współczynnik $m_{32}=\frac{1}{2}$. Teraz wierszem, którym się nie zmienia jest wiersz drugi!. Dlatego w mianowniku jest „2" a nie „-2". Postać nowego wiersza po eliminacji (wyraz pierwszy nie ulega zmianie – można to łatwo sprawdzić, bo stoi nad nim „0"):

$$a_{32}=1-\frac{1}{2}\cdot 2=1-1=0,\ a_{33}=9-\frac{1}{2}\cdot 1=9-\frac{1}{2}=\frac{17}{2},\ b_3=29-\frac{1}{2}\cdot 7=29-\frac{7}{2}=\frac{51}{2}.$$

Tablica wyrazów wygląda teraz następująco:

$$\begin{vmatrix} 1 & -2 & -1 & -6 \\ 0 & 2 & 1 & 7 \\ 0 & 0 & \frac{17}{2} & \frac{51}{2} \end{vmatrix}.$$

Postać macierzy górnotrójkątnej: $\boldsymbol{U} = \begin{bmatrix} 1 & -2 & -1 \\ 0 & 2 & 1 \\ 0 & 0 & \frac{17}{2} \end{bmatrix}$. Macierz dolnotrójkątną tworzy się następująco: $\boldsymbol{L} = \begin{bmatrix} 1 & 0 & 0 \\ m_{21} & 1 & 0 \\ m_{31} & m_{32} & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ -1 & \frac{1}{2} & 1 \end{bmatrix}$. Łatwo sprawdzić, iż $\boldsymbol{LU} = \boldsymbol{A}$.

Teraz eliminacji podlegają wyrazy nad przekątną, kolejno „1”, „-1” oraz „-2”. Do eliminacji „1” współczynnik $m_{23} = \frac{1}{17/2} = \frac{2}{17}$ a do eliminacji „-1”: $m_{13} = \frac{-1}{17/2} = -\frac{2}{17}$.

Postać tablicy po przekształceniach:

$$\begin{vmatrix} 1 & -2 & 0 & -3 \\ 0 & 2 & 0 & 4 \\ 0 & 0 & \frac{17}{2} & \frac{51}{2} \end{vmatrix}.$$

Ostatni krok wymaga wyzerowania „-2”. Ostatnie $m_{12} = \frac{-2}{2} = -1$.
Końcowa postać tablicy (macierz $\boldsymbol{A}$ jest teraz diagonalna):

$$\begin{vmatrix} 1 & 0 & 0 & 1 \\ 0 & 2 & 0 & 4 \\ 0 & 0 & \frac{17}{2} & \frac{51}{2} \end{vmatrix}.$$

Ostatnie przekształcenie polega na podzieleniu ostatniej kolumny wyrazów wolnych przez odpowiednie wyrazy diagonalne ($b_1 = \frac{1}{1} = 1, \quad b_2 = \frac{4}{2} = 2, \quad b_3 = \frac{51}{2} \cdot \frac{2}{17} = 3$). Z własności macierzy jednostkowej ($\begin{vmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 2 \\ 0 & 0 & 1 & 3 \end{vmatrix}$, $\boldsymbol{Ix} = \boldsymbol{b} \rightarrow \boldsymbol{x} = \boldsymbol{b}$) wynika, iż wyrazy wolne są poszukiwanymi rozwiązaniami wyjściowego układu, czyli:

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

Przykład 2

Rozwiązać metodą eliminacji Gaussa – Jordana układ równań $\begin{bmatrix} 6 & 2 & 2 & 4 \\ -1 & 2 & 2 & -3 \\ 0 & 1 & 1 & 4 \\ 1 & 0 & 2 & 3 \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 1 \\ -1 \\ 2 \\ 1 \end{bmatrix}$.

Na podstawie układu budujemy tablicę:

$$\left|\begin{array}{ccccc} 6 & 2 & 2 & 4 & 1 \\ -1 & 2 & 2 & -3 & -1 \\ 0 & 1 & 1 & 4 & 2 \\ 1 & 0 & 2 & 3 & 1 \end{array}\right| \rightarrow \left|\begin{array}{ccccc} 6 & 2 & 2 & 4 & 1 \\ 0 & \frac{7}{3} & \frac{7}{3} & -\frac{7}{3} & -\frac{5}{6} \\ 0 & 1 & 1 & 4 & 2 \\ 0 & -\frac{1}{3} & \frac{5}{3} & \frac{7}{3} & \frac{5}{6} \end{array}\right| \rightarrow \left|\begin{array}{ccccc} 6 & 2 & 2 & 4 & 1 \\ 0 & \frac{7}{3} & \frac{7}{3} & -\frac{7}{3} & -\frac{5}{6} \\ 0 & 0 & 0 & 5 & \frac{33}{14} \\ 0 & 0 & 2 & 2 & \frac{5}{7} \end{array}\right|.$$

Współczynniki do wyzerowania pierwszej kolumny: $m_{21} = -\frac{1}{6},\ m_{31} = 0,\ m_{41} = \frac{1}{6}$, do drugiej: $m_{32} = \frac{3}{7},\ m_{42} = -\frac{1}{7}$. Dalej konieczne jest wyzerowanie wyrazu $a_{43} = 2$. Jednak obliczenie współczynnika m_{43} wymagałoby dzielenia przez zero ($m_{43} = \frac{2}{"0"}$). Czy to oznacza, że wyjściowa macierz była osobliwa? Nie, po prostu wyjściowa kolejność równań powoduje takie ułożenie współczynników macierzy – w takim wypadku należy zamienić kolejność wierszy – w powyższym przykładzie ulegną zamianie wiersze trzeci i czwarty. Wtedy zero wskoczy na właściwe sobie miejsce. Natomiast osobliwość macierzy skutkowałaby w postaci późniejszego dzielenia przez zero w czasie obliczania wyrazów wektora rozwiązań.

Dalsze przekształcenia tablicy:

$$\left|\begin{array}{ccccc} 6 & 2 & 2 & 4 & 1 \\ 0 & \frac{7}{3} & \frac{7}{3} & -\frac{7}{3} & -\frac{5}{6} \\ 0 & 0 & 2 & 2 & \frac{5}{7} \\ 0 & 0 & 0 & 5 & \frac{33}{14} \end{array}\right| \rightarrow \left|\begin{array}{ccccc} 6 & 2 & 2 & 0 & -\frac{31}{35} \\ 0 & \frac{7}{3} & \frac{7}{3} & 0 & \frac{8}{15} \\ 0 & 0 & 2 & 0 & -\frac{8}{35} \\ 0 & 0 & 0 & 5 & \frac{33}{14} \end{array}\right| \rightarrow \left|\begin{array}{ccccc} 6 & 2 & 0 & 0 & -\frac{23}{35} \\ 0 & \frac{7}{3} & 0 & 0 & \frac{8}{15} \\ 0 & 0 & 2 & 0 & -\frac{8}{35} \\ 0 & 0 & 0 & 5 & \frac{33}{14} \end{array}\right|$$

$$\left|\begin{array}{cccc|c} 6 & 0 & 0 & 0 & -\frac{39}{35} \\ 0 & \frac{7}{3} & 0 & 0 & \frac{8}{15} \\ 0 & 0 & 2 & 0 & -\frac{8}{35} \\ 0 & 0 & 0 & 5 & \frac{33}{14} \end{array}\right| \rightarrow \left|\begin{array}{cccc|c} 1 & 0 & 0 & 0 & -\frac{13}{70} \\ 0 & 1 & 0 & 0 & \frac{8}{35} \\ 0 & 0 & 1 & 0 & -\frac{4}{35} \\ 0 & 0 & 0 & 1 & \frac{33}{70} \end{array}\right| \rightarrow \left|\begin{array}{l} x_1 = -\frac{13}{70} \\ x_2 = \frac{8}{35} \\ x_3 = -\frac{4}{35} \\ x_4 = \frac{33}{70} \end{array}\right| .$$

METODA CHOLESKIEGO

Metoda opracowana jest dla macierzy współczynników symetrycznych dodatnio określonych. Dzięki takim własnością macierzy $\boldsymbol{A}$ jest możliwy następujący jej rozkład: $\boldsymbol{A} = \boldsymbol{L} \cdot \boldsymbol{L}^{\boldsymbol{T}}$.
Ze sprawdzeniem symetrii macierzy nie ma na ogół problemów, musi być spełniony warunek: $\boldsymbol{A} = \boldsymbol{A}^T \quad \rightarrow \quad a_{ij} = a_{ji}, \quad i, j = 1, 2, ..., n$. Natomiast badanie dodatniej określoności jest na ogół kłopotliwe, dlatego pomocne może okazać się następujące twierdzenie:

Twierdzenie 1
Jeżeli macierz $\boldsymbol{A}$ o współczynnikach rzeczywistych jest symetryczna i ściśle dominująca na przekątnej głównej i dodatkowo posiada wszystkie elementy diagonalne dodatnie, to macierz $\boldsymbol{A}$ jest dodatnio określona.

Macierz nazywamy ściśle dominującą na przekątnej głównej, jeżeli: $|a_{ii}| > \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}|, \quad i = 1, 2, 3, ..., n$.

Wzór na rozkład macierzy w metodzie Choleskiego oraz wzory na niewiadome wektory $\boldsymbol{x}$ i $\boldsymbol{y}$:

$$\begin{cases} l_{jj} = \sqrt{a_{jj} - \sum_{k=1}^{j-1} l_{jk}^2} \\ l_{ij} = \frac{1}{l_{jj}} (a_{ij} - \sum_{k=1}^{j-1} l_{ik} \cdot l_{jk}) \end{cases} , \quad j = 1, 2, ..., n; \quad i = j+1, ..., n$$

$$y_i = \frac{1}{l_{ii}} (b_i - \sum_{j=1}^{i-1} l_{ij} \cdot y_j), \quad i = 1, ..., n \qquad .$$

$$x_i = \left[y_i - \sum_{j=i+1}^{n} l_{ji} \cdot x_j \right] \cdot \frac{1}{l_{ii}}, \quad i = n, ..., 1$$

Przykład 3

Rozwiązać układ równań $\boldsymbol{Ax}=\boldsymbol{b}$, gdzie $\boldsymbol{A}=\begin{bmatrix}4 & -2 & 0\\ -2 & 5 & -2\\ 0 & -2 & 5\end{bmatrix}$, $\boldsymbol{b}=\begin{bmatrix}-6\\ 3\\ 8\end{bmatrix}$ metodą eliminacji Choleskiego.

Aby zastosować metodę Choleskiego do nieosobliwego układu równań, należy sprawdzić warunek symetrii i dodatniej określoności macierzy $\boldsymbol{A}$. Symetria jest spełniona, gdyż $a_{12}=a_{21}=-2,\quad a_{13}=a_{31}=-1,\quad a_{23}=a_{32}=3.$ Do zbadania dodatniej określoności wykorzystamy tw.1: macierz symetryczna jest dominująca na przekątnej głównej, gdyż: $4>|-2|+|0|=2,\quad 5>|-2|+|-2|=4,\quad 5>|0|+|-2|=2$.

Rozłożenie macierzy $\boldsymbol{A}$ na czynniki trójkątne $\boldsymbol{L}\cdot\boldsymbol{L}^{\boldsymbol{T}}$:

$$\begin{bmatrix}4 & -2 & 0\\ -2 & 5 & -2\\ 0 & -2 & 5\end{bmatrix}=\begin{bmatrix}l_{11} & 0 & 0\\ l_{21} & l_{22} & 0\\ l_{31} & l_{32} & l_{33}\end{bmatrix}\cdot\begin{bmatrix}l_{11} & l_{21} & l_{31}\\ 0 & l_{22} & l_{32}\\ 0 & 0 & l_{33}\end{bmatrix}$$

Dokonując odpowiednich mnożeń wierszy macierzy $\boldsymbol{L}$ i kolumn macierzy $\boldsymbol{L}^{\boldsymbol{T}}$ i porównując wynik z odpowiednim wyrazem macierzy $\boldsymbol{A}$ wyznaczamy nieznane wyrazy l_{ij}:

$$l_{11}\cdot l_{11}=a_{11}=4\quad\rightarrow\quad l_{11}=\sqrt{4}=2$$

$$l_{11}\cdot l_{21}=a_{21}=-2\quad\rightarrow\quad l_{21}=\frac{-2}{l_{11}}=\frac{-2}{2}=-1$$

$$l_{11}\cdot l_{31}=a_{31}=0\quad\rightarrow\quad l_{31}=\frac{0}{l_{11}}=0$$

$$l_{21}^2+l_{22}^2=a_{22}=5\quad\rightarrow\quad l_{22}=\sqrt{5-l_{21}^2}=\sqrt{5-1}=2$$

$$l_{31}\cdot l_{21}+l_{32}\cdot l_{22}=a_{32}=-2\quad\rightarrow\quad l_{32}=\frac{-2-l_{31}\cdot l_{21}}{l_{22}}=\frac{-2-0}{2}=-1$$

$$l_{21}^2+l_{22}^2+l_{33}^2=a_{33}=5\quad\rightarrow\quad l_{33}=\sqrt{5-l_{31}^2-l_{32}^2}=\sqrt{5-0-1}=2$$

Macierz dolnotrójkątna: $\boldsymbol{L}=\begin{bmatrix}l_{11} & 0 & 0\\ l_{21} & l_{22} & 0\\ l_{31} & l_{32} & l_{33}\end{bmatrix}=\begin{bmatrix}2 & 0 & 0\\ -1 & 2 & 0\\ 0 & -1 & 2\end{bmatrix}$.

Macierz górnotrójkątna: $\boldsymbol{U}=\boldsymbol{L}^{T}==\begin{bmatrix}l_{11} & l_{21} & l_{31}\\ 0 & l_{22} & l_{32}\\ 0 & 0 & l_{33}\end{bmatrix}=\begin{bmatrix}2 & -1 & 0\\ 0 & 2 & -1\\ 0 & 0 & 2\end{bmatrix}$.

Krok wprzód $\boldsymbol{Ly}=\boldsymbol{b}$:

$$\begin{bmatrix} 2 & 0 & 0 \\ -1 & 2 & 0 \\ 0 & -1 & 2 \end{bmatrix} \cdot \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} -6 \\ 3 \\ 8 \end{bmatrix} \quad \rightarrow \quad \begin{array}{c} y_1 = \dfrac{-6}{2} = -3 \\ y_2 = \dfrac{1}{2}(3 + y_1) = 0 \\ y_3 = \dfrac{1}{2}(8 - y_2) = 4 \end{array} \quad \rightarrow \quad \boldsymbol{y} = \begin{bmatrix} -3 \\ 0 \\ 4 \end{bmatrix}$$

Krok wstecz $\boldsymbol{L}^T \boldsymbol{x} = \boldsymbol{y}$:

$$\begin{bmatrix} 2 & -1 & 0 \\ 0 & 2 & -1 \\ 0 & 0 & 2 \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} -3 \\ 0 \\ 4 \end{bmatrix} \quad \rightarrow \quad \begin{array}{c} x_3 = \dfrac{4}{2} = 2 \\ x_2 = \dfrac{1}{2}(0 + x_3) = 1 \\ x_3 = \dfrac{1}{2}(-3 + x_2) = -1 \end{array} \quad \rightarrow \quad \boldsymbol{x} = \begin{bmatrix} -1 \\ 1 \\ 2 \end{bmatrix}.$$

Ostatecznym wektorem rozwiązań jest wektor $\boldsymbol{x}$.

Wymaganie związane z dodatnią określonością macierzy $\boldsymbol{A}$ może być w wyjątkowych sytuacjach niespełnione. Wtedy macierze trójkątne $\boldsymbol{L} \cdot \boldsymbol{L}^T$ istnieją w dziedzinie liczb zespolonych, ale końcowe rozwiązanie jest rzeczywiste o ile tylko układ nie jest osobliwy.

Metody iteracyjne, w odróżnieniu od metod eliminacyjnych, dostarczają w wyniku metod iteracji prostej (z relaksacją) całego zbioru przybliżeń wektora rozwiązań, który przy odpowiedniej liczbie iteracji będzie zbieżny do rozwiązania ścisłego $\boldsymbol{x} = \boldsymbol{A}^{-1}\boldsymbol{b}$.
W metodach iteracyjnych z każdego z równań wyznaczamy niewiadomą (z „i"-tego równania pochodzi „i-ta" niewiadoma) za pomocą wszystkich pozostałych. Niewiadome te podlegają obliczeniu na podstawie znajomości poprzedniego przybliżenia, na samym początku na znajomości wektora startowego. Tak działa metoda Jacobiego, która zawsze korzysta z wyników z poprzedniej iteracji, natomiast metoda Gaussa – Seidela korzysta z informacji z aktualnej iteracji, jeżeli jest to już możliwe. Metody te są zbieżne, jeżeli macierz $\boldsymbol{A}$ jest dodatnio określona (jest to warunek wystarczający zbieżności).

Sformułowanie problemu: $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}, \quad \det(\boldsymbol{A}) \neq 0 \quad \rightarrow \quad \sum_{j=1}^{n} a_{ij} x_j = b_i, \quad i = 1, 2, ..., n$.

METODA JACOBIEGO

$$\begin{cases} \boldsymbol{x}^{(0)} = \{x_1^{(0)}, x_2^{(0)}, ..., x_n^{(0)}\} \\ x_i^{(k+1)} = \dfrac{1}{a_{ii}}(b_i - \sum\limits_{\substack{j=1 \\ j \neq i}}^{n} a_{ij} \cdot x_j^{(k)}), \quad i = 1, 2, ..., n \end{cases}$$

Po rozłożeniu macierzy $\boldsymbol{A}$ na składniki: $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b} \quad \rightarrow \quad \boldsymbol{L}\boldsymbol{x} + \boldsymbol{D}\boldsymbol{x} + \boldsymbol{U}\boldsymbol{x} = \boldsymbol{b}$ można algorytm sformułować w zapisie macierzowym:

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

METODA GAUSSA - SEIDELA

$$\begin{cases} \boldsymbol{x}^{(0)} = \{x_1^{(0)}, x_2^{(0)}, ..., x_n^{(0)}\} \\ x_i^{(k+1)} = \dfrac{1}{a_{ii}}(b_i - \sum\limits_{j=1}^{i-1} a_{ij} \cdot x_j^{(k+1)} - \sum\limits_{j=i+1}^{n} a_{ij} \cdot x_j^{(k)}), \quad i = 1, 2, ..., n \end{cases}$$

W zapisie macierzowym:

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

Kryteria przerwania procesu iteracyjnego są takie same dla obydwu powyższych metod:

1. kontrola tempa zbieżności: $\varepsilon^{(1)} = \dfrac{\left\| \boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)} \right\|}{\left\| \boldsymbol{x}^{(k+1)} \right\|} \le \varepsilon_{dop}^{(1)}$,
2. kontrola wielkości residuum: $\varepsilon^{(2)} = \dfrac{\left\| \boldsymbol{A} \cdot \boldsymbol{x}^{(k+1)} - \boldsymbol{b} \right\|}{\left\| \boldsymbol{A} \cdot \boldsymbol{x}_0 - \boldsymbol{b} \right\|} \le \varepsilon_{dop}^{(2)}$.

Zastosowanie relaksacji polega na poprawieniu wektora rozwiązań po każdym kroku iteracyjnym wg wzoru:

$$\tilde{x}_i^{(k+1)} = x_i^{(k)} + \lambda \cdot (x_i^{(k+1)} - x_i^{(k)}),$$

gdzie λ jest parametrem relaksacji (przyjmowanym arbitralnie na początku lub ustalanym dynamicznie po każdym kroku).

Przykład 4

Rozwiązać układ równań $\boldsymbol{Ax} = \boldsymbol{b}$, gdzie $\boldsymbol{A} = \begin{bmatrix} 4 & -2 & 0 \\ -2 & 5 & -2 \\ 0 & -2 & 5 \end{bmatrix}$, $\boldsymbol{b} = \begin{bmatrix} -6 \\ 3 \\ 8 \end{bmatrix}$ metodą iteracji Jacobiego. Przyjąć wektor startowy $\boldsymbol{x}^{(0)} = \{0, 0, 0\}$.

Po zapisaniu tradycyjnym powyższego układu i wyliczeniu z każdego z równań kolejnych niewiadomych, otrzymujemy schemat iteracyjny metody Jacobiego.

$$\begin{cases} 4x_1 - 2x_2 = -6 \\ -2x_1 + 5x_2 - 2x_3 = 3 \\ -2x_2 + 5x_3 = 8 \end{cases} \rightarrow \begin{cases} x_1^{(k+1)} = \dfrac{1}{4}(-6 + 2x_2^{(k)}) \\ x_2^{(k+1)} = \dfrac{1}{5}(3 + 2x_1^{(k)} + 2x_3^{(k)}) \\ x_3^{(k+1)} = \dfrac{1}{5}(8 + 2x_2^{(k)}) \end{cases}.$$

Rozpoczynając obliczenia od wektora startowego $\boldsymbol{x}^{(0)} = \{0,0,0\}$ otrzymujemy ciąg przybliżeń wektora rozwiązań, po każdym kroku kontrolując błąd obliczeń (tempo zbieżności i residuum liczone dla dwóch rodzajów norm: euklidesowej i maksimum):

Iteracja pierwsza ($k = 0$):

$$\begin{cases} x_1^{(1)} = \frac{1}{4}(-6 + 2x_2^{(0)}) = \frac{1}{4}(-6+0) = -1.5 \\ x_2^{(1)} = \frac{1}{5}(3 + 2x_1^{(0)} + 2x_3^{(0)}) = \frac{1}{5}(3+0+0) = 0.6 \\ x_3^{(1)} = \frac{1}{5}(8 + 2x_2^{(0)}) = \frac{1}{5}(8+0) = 1.6 \end{cases}$$

$$\boldsymbol{\varepsilon}^{(1)} = \frac{\left\|\boldsymbol{x}^{(1)} - \boldsymbol{x}^{(0)}\right\|}{\left\|\boldsymbol{x}^{(1)}\right\|}, \quad \begin{cases} \varepsilon_e^{(1)} = 1.0 \\ \varepsilon_m^{(1)} = 1.0 \end{cases}, \quad \boldsymbol{\varepsilon}^{(2)} = \frac{\left\|\boldsymbol{A}\boldsymbol{x}^{(1)} - \boldsymbol{b}\right\|}{\left\|\boldsymbol{A}\boldsymbol{x}^{(0)} - \boldsymbol{b}\right\|}, \quad \begin{cases} \varepsilon_e^{(2)} = 0.163673 \\ \varepsilon_m^{(2)} = 0.150 \end{cases}$$

Iteracja druga ($k = 1$):

$$\begin{cases} x_1^{(2)} = \frac{1}{4}(-6 + 2x_2^{(1)}) = \frac{1}{4}(-6 + 2\cdot 0.6) = -1.2 \\ x_2^{(2)} = \frac{1}{5}(3 + 2x_1^{(1)} + 2x_3^{(1)}) = \frac{1}{5}(3 + 2\cdot(-1.5) + 2\cdot 1.6) = 0.64 \\ x_3^{(2)} = \frac{1}{5}(8 + 2x_2^{(1)}) = \frac{1}{5}(8 + 2\cdot 0.6) = 1.84 \end{cases}$$

$$\boldsymbol{\varepsilon}^{(1)} = \frac{\left\|\boldsymbol{x}_2 - \boldsymbol{x}_1\right\|}{\left\|\boldsymbol{x}_2\right\|}, \quad \begin{cases} \varepsilon_e^{(1)} = 0.1019 \\ \varepsilon_m^{(1)} = 0.1630 \end{cases}, \quad \boldsymbol{\varepsilon}^{(2)} = \frac{\left\|\boldsymbol{A}\boldsymbol{x}_2 - \boldsymbol{b}\right\|}{\left\|\boldsymbol{A}\boldsymbol{x}_0 - \boldsymbol{b}\right\|}, \quad \begin{cases} \varepsilon_e^{(2)} = 0.1040 \\ \varepsilon_m^{(2)} = 0.1350 \end{cases}$$

Iteracja trzecia ($k = 2$):

$$\begin{cases} x_1^{(3)} = \frac{1}{4}(-6 + 2x_2^{(2)}) = \frac{1}{4}(-6 + 2\cdot 0.64) = -1.18 \\ x_2^{(3)} = \frac{1}{5}(3 + 2x_1^{(2)} + 2x_3^{(2)}) = \frac{1}{5}(3 + 2\cdot(-1.2) + 2\cdot 1.84) = 0.856 \\ x_3^{(3)} = \frac{1}{5}(8 + 2x_2^{(2)}) = \frac{1}{5}(8 + 2\cdot 0.64) = 1.856 \end{cases}$$

$$\boldsymbol{\varepsilon}^{(1)} = \frac{\left\|\boldsymbol{x}_3 - \boldsymbol{x}_2\right\|}{\left\|\boldsymbol{x}_3\right\|}, \quad \begin{cases} \varepsilon_e^{(1)} = 0.0922 \\ \varepsilon_m^{(1)} = 0.1164 \end{cases}, \quad \boldsymbol{\varepsilon}^{(2)} = \frac{\left\|\boldsymbol{A}\boldsymbol{x}_3 - \boldsymbol{b}\right\|}{\left\|\boldsymbol{A}\boldsymbol{x}_0 - \boldsymbol{b}\right\|}, \quad \begin{cases} \varepsilon_e^{(2)} = 0.0589 \\ \varepsilon_m^{(2)} = 0.0540 \end{cases}$$

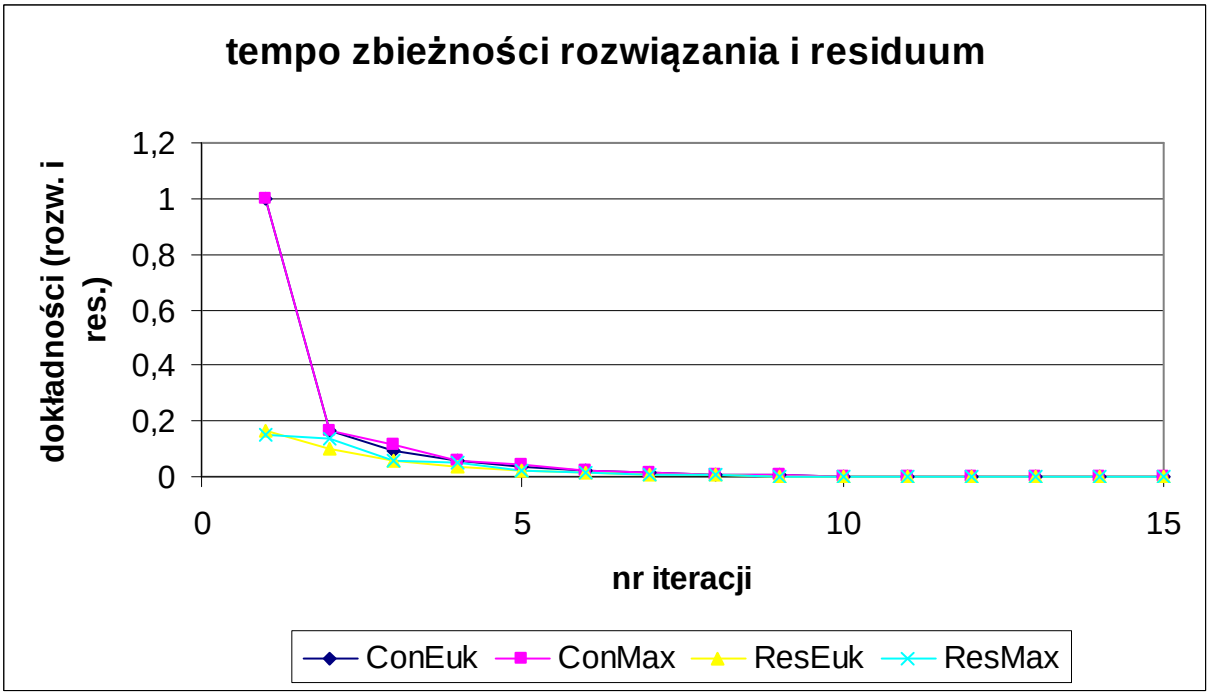
Proces jest bardzo wolno zbieżny do rozwiązania ścisłego $\bar{\boldsymbol{x}} = \{-1,1,2\}$. Po piętnastu iteracjach otrzymano wynik $\boldsymbol{x}^{(15)} = \{$-1.000392, 0.999687, 1.999687$\}$. Aby przyśpieszyć obliczenia, można zastosować technikę, np. nadrelaksacji z parametrem $\lambda = 1.6$. Wtedy poprawione rozwiązania po drugim kroku iteracji wynosić będą:

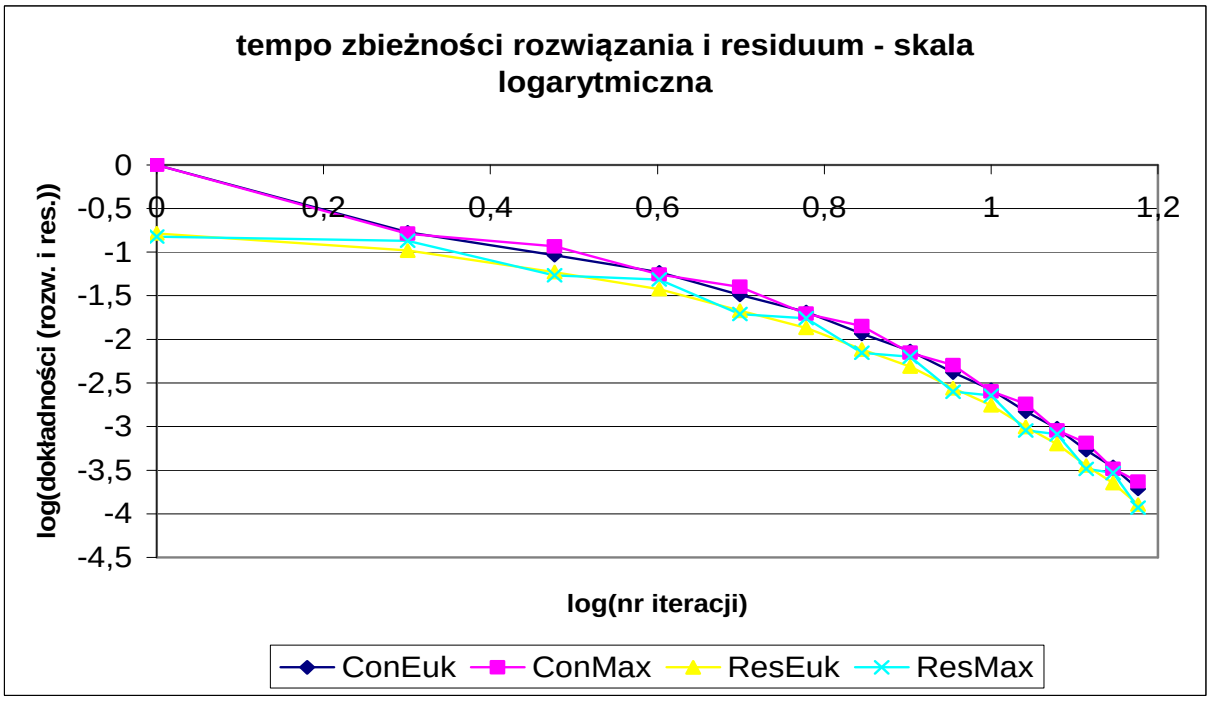
$$\begin{cases} \bar{x}_1^{(2)} = x_1^{(1)} + \lambda \cdot (x_1^{(2)} - x_1^{(1)}) = -1.5 + 1.6 \cdot (-1.2 + 1.5) = -1.02 \\ \bar{x}_2^{(2)} = x_2^{(1)} + \lambda \cdot (x_2^{(2)} - x_2^{(1)}) = 0.6 + 1.6 \cdot (0.64 - 0.6) = 0.664 \\ \bar{x}_3^{(2)} = x_3^{(1)} + \lambda \cdot (x_3^{(2)} - x_3^{(1)}) = 1.6 + 1.6 \cdot (1.84 - 1.6) = 1.984 \end{cases}$$

Dopiero dla tych wyników policzone błędy wynoszą:

$$\varepsilon^{(1)} = \frac{\|\bar{\boldsymbol{x}}_2 - \boldsymbol{x}_1\|}{\|\bar{\boldsymbol{x}}_2\|}, \quad \begin{cases} \varepsilon_e^{(1)} = 0.1019 \\ \varepsilon_m^{(1)} = 0.2419 \end{cases}, \quad \varepsilon^{(2)} = \frac{\|\boldsymbol{A}\bar{\boldsymbol{x}}_2 - \boldsymbol{b}\|}{\|\boldsymbol{A}\boldsymbol{x}_0 - \boldsymbol{b}\|}, \quad \begin{cases} \varepsilon_e^{(2)} = 0.1736 \\ \varepsilon_m^{(2)} = 0.2010 \end{cases}$$

Poniższe wykresy przedstawiają tempa zbieżności rozwiązania i residuum równania dla opcji metody bez relaksacji w normach: dziesiętnej i logarytmicznej.





Można spodziewać się większego przyśpieszenia zbieżności po zastosowaniu metody iteracyjnej Gaussa – Seidela.

Przykład 5
Rozwiązać powyższe zadanie metodą iteracji Gaussa – Seidela.

Wyjściowy układ równań: $\boldsymbol{Ax}=\boldsymbol{b}$, gdzie $\boldsymbol{A}=\begin{bmatrix}4 & -2 & 0\\ -2 & 5 & -2\\ 0 & -2 & 5\end{bmatrix}$, $\boldsymbol{b}=\begin{bmatrix}-6\\ 3\\ 8\end{bmatrix}$.

Schemat iteracyjny metody Gaussa – Seidela (zmodyfikowany schemat metody Jacobiego):

$$\begin{cases}4x_1-2x_2=-6\\ -2x_1+5x_2-2x_3=3\\ -2x_2+5x_3=8\end{cases}\rightarrow\begin{cases}x_1^{(k+1)}=\frac{1}{4}(-6+2x_2^{(k)})\\ x_2^{(k+1)}=\frac{1}{5}(3+2x_1^{(k+1)}+2x_3^{(k)})\\ x_3^{(k+1)}=\frac{1}{5}(8+2x_2^{(k+1)})\end{cases}.$$

Tam gdzie to jest możliwe wykorzystuje się już informację „najświeższą” z aktualnego kroku iteracyjnego $k+1$.

Iteracja pierwsza ($k=0$):

$$\begin{cases} x_1^{(1)} = \frac{1}{4}(-6 + 2x_2^{(0)}) = \frac{1}{4}(-6+0) = -1.5 \\ x_2^{(1)} = \frac{1}{5}(3 + 2x_1^{(1)} + 2x_3^{(0)}) = \frac{1}{5}(3 + 2\cdot(-1.5) + 0) = 0.0 \\ x_3^{(1)} = \frac{1}{5}(8 + 2x_2^{(1)}) = \frac{1}{5}(8+0) = 1.6 \end{cases}$$

$$\varepsilon^{(1)} = \frac{\left\| \boldsymbol{x}^{(1)} - \boldsymbol{x}^{(0)} \right\|}{\left\| \boldsymbol{x}^{(1)} \right\|}, \quad \begin{cases} \varepsilon_e^{(1)} = 1.0 \\ \varepsilon_m^{(1)} = 1.0 \end{cases}, \quad \varepsilon^{(2)} = \frac{\left\| \boldsymbol{A}\boldsymbol{x}^{(1)} - \boldsymbol{b} \right\|}{\left\| \boldsymbol{A}\boldsymbol{x}^{(0)} - \boldsymbol{b} \right\|}, \quad \begin{cases} \varepsilon_e^{(2)} = 0.3065 \\ \varepsilon_m^{(2)} = 0.4000 \end{cases}$$

Iteracja druga ($k = 1$):

$$\begin{cases} x_1^{(2)} = \frac{1}{4}(-6 + 2x_2^{(1)}) = \frac{1}{4}(-6 + 2\cdot 0.0) = -1.50 \\ x_2^{(2)} = \frac{1}{5}(3 + 2x_1^{(2)} + 2x_3^{(1)}) = \frac{1}{5}(3 + 2\cdot(-1.50) + 2\cdot 1.6) = 0.640 \\ x_3^{(2)} = \frac{1}{5}(8 + 2x_2^{(2)}) = \frac{1}{5}(8 + 2\cdot 0.64) = 1.8560 \end{cases}$$

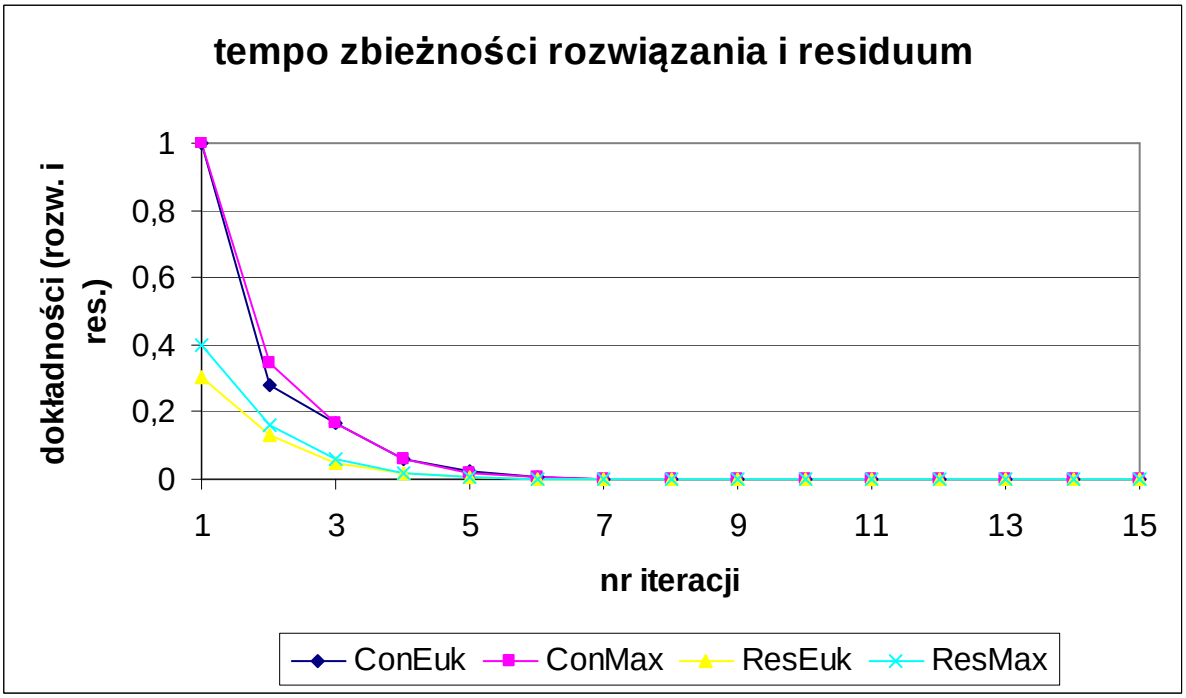
$$\varepsilon^{(1)} = \frac{\left\| \boldsymbol{x}_2 - \boldsymbol{x}_1 \right\|}{\left\| \boldsymbol{x}_2 \right\|}, \quad \begin{cases} \varepsilon_e^{(1)} = 0.2790 \\ \varepsilon_m^{(1)} = 0.3448 \end{cases}, \quad \varepsilon^{(2)} = \frac{\left\| \boldsymbol{A}\boldsymbol{x}_2 - \boldsymbol{b} \right\|}{\left\| \boldsymbol{A}\boldsymbol{x}_0 - \boldsymbol{b} \right\|}, \quad \begin{cases} \varepsilon_e^{(2)} = 0.1320 \\ \varepsilon_m^{(2)} = 0.1600 \end{cases}$$

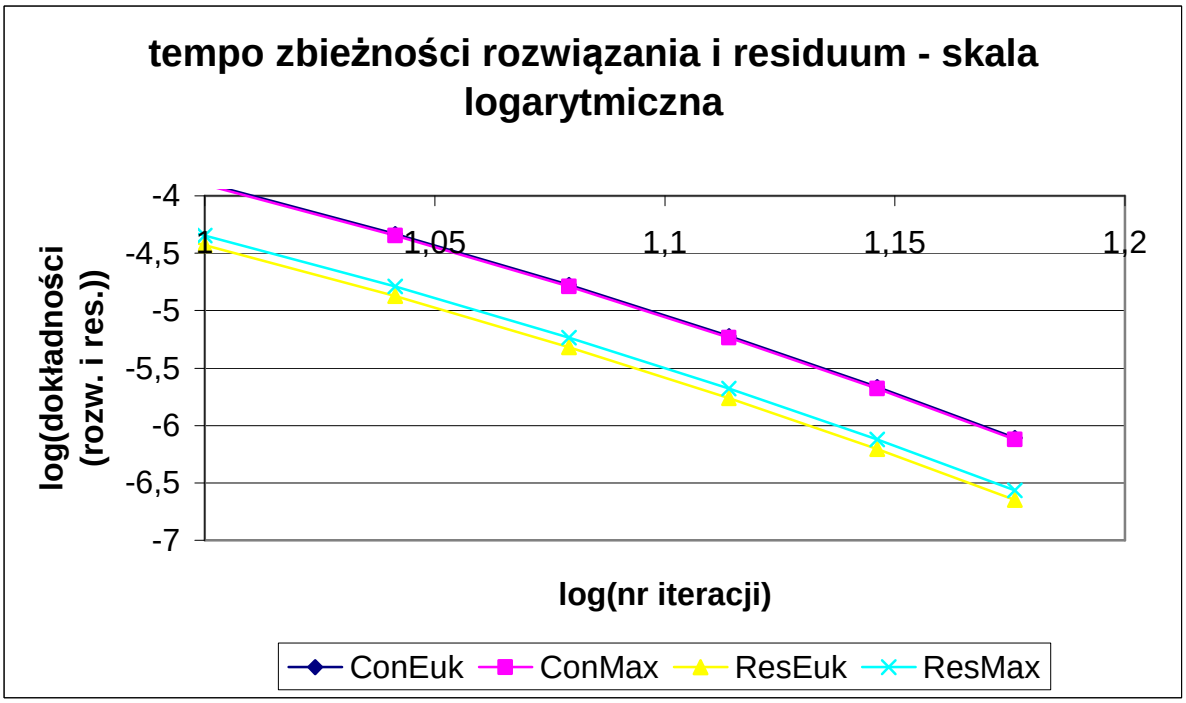
Iteracja trzecia ($k = 2$):

$$\begin{cases} x_1^{(3)} = \frac{1}{4}(-6 + 2x_2^{(2)}) = \frac{1}{4}(-6 + 2\cdot 0.640) = -1.18 \\ x_2^{(3)} = \frac{1}{5}(3 + 2x_1^{(3)} + 2x_3^{(2)}) = \frac{1}{5}(3 + 2\cdot(-1.18) + 2\cdot 1.856) = 0.8704 \\ x_3^{(3)} = \frac{1}{5}(8 + 2x_2^{(3)}) = \frac{1}{5}(8 + 2\cdot 0.8704) = 1.9482 \end{cases}$$

$$\varepsilon^{(1)} = \frac{\left\| \boldsymbol{x}_3 - \boldsymbol{x}_2 \right\|}{\left\| \boldsymbol{x}_3 \right\|}, \quad \begin{cases} \varepsilon_e^{(1)} = 0.1661 \\ \varepsilon_m^{(1)} = 0.1643 \end{cases}, \quad \varepsilon^{(2)} = \frac{\left\| \boldsymbol{A}\boldsymbol{x}_3 - \boldsymbol{b} \right\|}{\left\| \boldsymbol{A}\boldsymbol{x}_0 - \boldsymbol{b} \right\|}, \quad \begin{cases} \varepsilon_e^{(2)} = 0.0475 \\ \varepsilon_m^{(2)} = 0.0576 \end{cases}$$

Po piętnastu iteracjach otrzymano rozwiązanie $\boldsymbol{x}^{(15)} = \{-1.0000,\ 1.0000,\ 2.0000\}$ z dokładnością do sześciu miejsc po przecinku. Wykresy zbieżności przedstawiono poniżej.


tempo zbieżności rozwiązania i residuum
dokładności (rozw. i res.)
1
0,8
0,6
0,4
0,2
0
1
3
5
7
9
11
13
15
nr iteracji
ConEuk
ConMax
ResEuk
ResMax


tempo zbieżności rozwiązania i residuum - skala logarytmiczna
log(dokładności (rozw. i res.))
-4
-4,5
-5
-5,5
-6
-6,5
-7
1
1,05
1,1
1,15
1,2
log(nr iteracji)
ConEuk
ConMax
ResEuk
ResMax

ODWRACANIE MACIERZY

Odwracanie macierzy dolnotrójkątnej

Dana jest macierz dolnotrójkątna $\boldsymbol{L}$ o wymiarze *n*, szukana jest macierz $\boldsymbol{C}$ taka, że $\boldsymbol{L}\cdot\boldsymbol{C}=\boldsymbol{I}$. Macierz $\boldsymbol{C}$, odwrotna do macierzy $\boldsymbol{L}$ jest również macierzą dolnotrójkątna.

Wzory ogólne:

$$\begin{cases} c_{ii}=\dfrac{1}{l_{ii}} \quad i=1,2,...,n \\ c_{ij}=-\dfrac{1}{l_{ii}}\sum_{k=j}^{i-1} l_{ik}\cdot c_{kj} \quad i=1,2,...,n \qquad j=1,2,...,i-1 \end{cases}$$

Przykład 13
Odwrócić macierz dolnotrójkątna.

$$\boldsymbol{L}=\begin{bmatrix}1&0&0\\2&4&0\\3&5&6\end{bmatrix} \qquad \boldsymbol{C}=\begin{bmatrix}c_{11}&0&0\\c_{21}&c_{22}&0\\c_{31}&c_{32}&c_{33}\end{bmatrix}$$

$$\boldsymbol{L}\cdot\boldsymbol{C}=\boldsymbol{I} \;\Rightarrow\; \begin{bmatrix}1&0&0\\2&4&0\\3&5&6\end{bmatrix}\cdot\begin{bmatrix}c_{11}&0&0\\c_{21}&c_{22}&0\\c_{31}&c_{32}&c_{33}\end{bmatrix}=\begin{bmatrix}1&0&0\\0&1&0\\0&0&1\end{bmatrix}$$

Dokonujemy mnożeń odpowiednich wierszy macierzy $\boldsymbol{L}$ i kolumn macierzy $\boldsymbol{C}$ tak, aby wyznaczyć wyrazy macierzy $\boldsymbol{C}$ za każdym razem porównując wyniki tych mnożeń z odpowiednim wyrazem macierzy jednostkowej.

$c_{11}\cdot 1+c_{21}\cdot 0+c_{31}\cdot 0=1 \;\rightarrow\; c_{11}=1$

$c_{11}\cdot 2+c_{21}\cdot 4+c_{31}\cdot 6=0 \;\rightarrow\; c_{21}=-\dfrac{1}{2}$

$c_{11}\cdot 3+c_{21}\cdot 5+c_{31}\cdot 0=0 \;\rightarrow\; c_{31}=-\dfrac{1}{12}$

$0\cdot 2+c_{22}\cdot 4+c_{32}\cdot 6=1 \;\rightarrow\; c_{22}=\dfrac{1}{4}$

$3\cdot 0+c_{22}\cdot 5+c_{32}\cdot 6=0 \;\rightarrow\; c_{32}=-\dfrac{5}{24}$

$0\cdot 3+0\cdot 5+c_{33}\cdot 0=1 \;\rightarrow\; c_{33}=\dfrac{1}{6}$

$$\boldsymbol{C} = \begin{bmatrix} 1 & 0 & 0 \\ -\frac{1}{2} & \frac{1}{4} & 0 \\ -\frac{1}{12} & -\frac{5}{24} & \frac{1}{6} \end{bmatrix}$$

Odwracanie macierzy górnotrójkątnej

Dana jest macierz górnotrójkątna $\boldsymbol{U}$ o wymiarze n, szukana jest macierz $\boldsymbol{C}$ taka, że $\boldsymbol{U} \cdot \boldsymbol{C} = \boldsymbol{I}$. Macierz $\boldsymbol{C}$, odwrotna do macierzy $\boldsymbol{U}$ jest również macierzą górnotrójkątna.

Wzory ogólne:

$$\begin{cases} c_{ii} = \dfrac{1}{u_{ii}} \quad i = n,...,1 \\ c_{ij} = -\dfrac{1}{u_{ii}} \sum_{k=i+1}^{j} u_{ik} \cdot c_{kj} \quad i = n,...,1 \qquad j = n,...,i+1 \end{cases}$$

Przykład 14
Odwrócić macierz górnotrójkątna.

$$\boldsymbol{U} = \begin{bmatrix} 1 & 2 & 3 \\ 0 & 4 & 5 \\ 0 & 0 & 6 \end{bmatrix} \qquad \boldsymbol{C} = \begin{bmatrix} c_{11} & c_{12} & c_{13} \\ 0 & c_{22} & c_{23} \\ 0 & 0 & c_{33} \end{bmatrix}$$

$$\boldsymbol{U} \cdot \boldsymbol{C} = \boldsymbol{I} \quad \Rightarrow \quad \begin{bmatrix} 1 & 2 & 3 \\ 0 & 4 & 5 \\ 0 & 0 & 6 \end{bmatrix} \cdot \begin{bmatrix} c_{11} & c_{12} & c_{13} \\ 0 & c_{22} & c_{23} \\ 0 & 0 & c_{33} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Postępowanie jest identyczne jak w przypadku macierzy dolnotrójkątnej.

$$c_{11}\cdot 1+2\cdot 0+3\cdot 0=1 \quad \rightarrow \quad c_{11}=1$$

$$c_{12}\cdot 0+c_{22}\cdot 4+5\cdot 0=0 \quad \rightarrow \quad c_{22}=\frac{1}{4}$$

$$c_{13}\cdot 0+c_{23}\cdot 0+c_{33}\cdot 6=0 \quad \rightarrow \quad c_{33}=\frac{1}{6}$$

$$c_{12}\cdot 1+c_{22}\cdot 2+3\cdot 0=1 \quad \rightarrow \quad c_{12}=-\frac{1}{2}$$

$$c_{13}\cdot 0+c_{23}\cdot 4+c_{33}\cdot 5=0 \quad \rightarrow \quad c_{23}=-\frac{5}{24}$$

$$c_{13}\cdot 1+c_{23}\cdot 2+c_{33}\cdot 3=1 \quad \rightarrow \quad c_{13}=-\frac{1}{12}$$

$$\boldsymbol{C}=\begin{bmatrix} 1 & -\frac{1}{2} & -\frac{1}{12} \\ 0 & \frac{1}{4} & -\frac{5}{24} \\ 0 & 0 & \frac{1}{6} \end{bmatrix}$$

Metoda Choleskiego

Jest to metoda odwracania macierzy symetrycznych, dodatnio określonych. Polega ona na rozłożeniu wyjściowej macierzy na czynniki trójkątne: $\boldsymbol{A} = \boldsymbol{L}\,\boldsymbol{L}^T$ a następnie na odwróceniu każdego z nich osobno i wymnożeniu tak, że: $\boldsymbol{A}^{-1} = \boldsymbol{L}^{-T}\,\boldsymbol{L}^{-1}$.

Wzory na rozkład macierzy $\boldsymbol{A}$ na czynniki trójkątne:

$$\begin{cases} l_{jj}=\sqrt{a_{jj}-\sum_{k=1}^{j-1} l_{jk}^2} & j=1,\ldots,n \\ l_{ij}=\frac{1}{l_{jj}}\left(a_{ij}-\sum_{k=1}^{j-1} l_{ik}\cdot l_{jk}\right) & i=j+1,\ldots,n \end{cases}$$

Po uzyskaniu macierzy dolnotrójkątnej $\boldsymbol{L}$ i górnotrójkątnej $\boldsymbol{L}^T$ odwraca się je korzystając ze wzorów zaprezentowanych w poprzednich podrozdziałach, a następnie mnoży obydwie macierze odwrotne, ale w odwrotnej kolejności.

Metody powiązane z rozwiązywaniem układów równań

Z definicji macierzy odwrotnej do macierzy $\boldsymbol{A}$ wynika następująca zależność:

$$\boldsymbol{A}\cdot\boldsymbol{A}^{-1}=\boldsymbol{I} \quad \Rightarrow \quad \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix} \cdot \begin{bmatrix} c_{11} & c_{12} & \dots & c_{1n} \\ c_{21} & c_{22} & \dots & c_{2n} \\ \dots & \dots & \dots & \dots \\ c_{n1} & c_{n2} & \dots & c_{nn} \end{bmatrix} = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 \end{bmatrix}$$

Powyższy zapis można rozbić na *n* układów równań, z których każdy służy do obliczenia kolejnej, *k-tej* kolumny macierzy $\boldsymbol{A}^{-1}$.

$$\begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix} \cdot \begin{bmatrix} c_{1k} \\ c_{2k} \\ \dots \\ c_{nk} \end{bmatrix} = \begin{bmatrix} b_{1k} \\ b_{2k} \\ \dots \\ b_{nk} \end{bmatrix} \quad k=1,2,...,n\,,$$

gdzie wyrazy wektora prawej strony: $b_{jk}=\begin{cases} 0 & dla \quad j\neq k \\ 1 & dla \quad j=k \end{cases}$.

W zależności od metody rozwiązywania tych układów równań można mówić o metodach eliminacji (np. *metoda eliminacji Gaussa* – wtedy rozwiązuje się jeden układ, ale z *n* prawymi stronami) lub metodach iteracyjnych (np. *metoda Jacobiego* lub *metoda Gaussa-Seidla*).

Metoda eliminacji Gaussa

Transformacji podlegają wyjściowa macierz nieosobliwa A oraz macierz C, która na początku obliczeń jest macierzą jednostkową, tzn. $C_{ij}=\begin{cases} 1, & i=j \\ 0, & i\neq j \end{cases}$.

$$\begin{aligned} a_{ij}^{(k)} &= a_{ij}^{(k-1)} - m_{ik}\cdot a_{kj}^{(k-1)} \\ c_{ij}^{(k)} &= c_{ij}^{(k-1)} - m_{ik}\cdot c_{kj}^{(k-1)} \end{aligned}, \text{ gdzie: } m_{ik}=\frac{a_{ik}^{(k-1)}}{a_{kk}^{(k-1)}}, \quad k=1,2,...,n-1; \quad i=k+1,...,n; \quad j=1,...,n$$

$$\begin{aligned} a_{ij}^{(k)} &= a_{ij}^{(k-1)} - m_{ik}\cdot a_{kj}^{(k-1)} \\ c_{ij}^{(k)} &= c_{ij}^{(k-1)} - m_{ik}\cdot c_{kj}^{(k-1)} \end{aligned}, \text{ gdzie: } m_{ik}=\frac{a_{ik}^{(k-1)}}{a_{kk}^{(k-1)}}, \quad k=n,n-1,...,2; \quad i=k-1,...,1; \quad j=1,...,n$$

$$c_{ij}=\frac{c_{ij}}{a_{ii}}, \quad i,j=1,2,...,n$$

NADOKREŚLONY UKŁAD RÓWNAŃ

Jeżeli w danym układzie równań liniowych $\boldsymbol{A}x = b$ jest więcej równań niż niewiadomych zmiennych, to taki układ nazywa się *nadokreślonym.* Jeżeli wszystkie równania są liniowo niezależne, to układ nie ma jednego wspólnego rozwiązania, tj. punktu, w którym wszystkie proste przecinają się.
W takim wypadku szuka się tzw. *pseudorozwiązania*, czyli punktu, który nie leży na żadnej prostej, ale jego odległości od każdej z prostych są minimalne w sensie jakieś normy.

Niech $\boldsymbol{A}$ będzie macierzą $n \times m$, gdzie *n* (liczba wierszy) oznacza liczbę równań, natomiast *m* (liczba kolumn) oznacza liczbę niewiadomych. W *układzie nadokreślonym: n>m*, w *układzie niedookreślonym: n<m*. Niech *b* będzie wektorem wyrazów wolnych o wymiarze *n*.
W pierwszym kroku buduje się wektor $\varepsilon = (\varepsilon_1, \varepsilon_2, ..., \varepsilon_n)$ zawierający odległości prostych od *pseudorozwiązania*. Następnie szuka się $\min \|\varepsilon\|$. Jeżeli zastosujemy normę średniokwadratową: $\|\varepsilon\| = \sqrt{\varepsilon_1^2 + \varepsilon_2^2 + ... + \varepsilon_n^2}$, to dalsze postępowanie nazywa się *metodą najmniejszych kwadratów*. Można też stosować normę maksimum.

Metoda najmniejszych kwadratów

Zapis wskaźnikowy (korzystny przy obliczeniach ręcznych):

$B = \sum_{i=1}^{n} (\sum_{j=1}^{m} a_{ij} x_j - b_i)^2$ - funkcjonał błędu

$\frac{\partial B}{\partial x_k} = 2\sum_{i=1}^{n} a_{ik} (\sum_{j=1}^{m} a_{ij} x_j - b_i) = 0$ - minimalizacja funkcjonału

Nowy układ równań liniowych (wymiar: $m \times m$):

$$\sum_{i=1}^{n} a_{ik} \sum_{j=1}^{m} a_{ij} x_j = \sum_{i=1}^{n} a_{ik} b_i, \quad k = 1, 2, ..., m$$

Zapis macierzowy (korzystny przy implementacji komputerowej):

$$\boldsymbol{B} = (\boldsymbol{A}x - b) \cdot (\boldsymbol{A}x - b)^T$$
$$\frac{\partial \boldsymbol{B}}{\partial x} = 2\boldsymbol{A}^{\boldsymbol{T}} (\boldsymbol{A}x - b) = 0$$
$$\boldsymbol{A}^T \boldsymbol{A}\, x = \boldsymbol{A}^T b$$

Przykład 11
Rozwiązać nadokreślony układ równań.

$$\begin{cases} x + y = 2 \\ x - y = 0 \\ x - 2y = -2 \end{cases} \Rightarrow \begin{cases} x + y - 2 = 0 \\ x - y = 0 \\ x - 2y + 2 = 0 \end{cases}$$

$$B(x,y) = \left\|\varepsilon(x,y)\right\|^2 = (x+y-2)^2 + (x-y)^2 + (x-2y+2)^2$$

$$\begin{cases} \dfrac{\partial B}{\partial x} = 2\cdot(x+y-2) + 2\cdot(x-y) + 2\cdot(x-2y+2) = 0 \\ \dfrac{\partial B}{\partial y} = 2\cdot(x+y-2) - 2\cdot(x-y) - 4\cdot(x-2y+2) = 0 \end{cases}$$

$$\begin{cases} 3x-2y=0 \\ -2x+6y=6 \end{cases} \Rightarrow \begin{cases} \boldsymbol{x_0 = \frac{6}{7} \approx 0.857143} \\ \boldsymbol{y_0 = \frac{9}{7} \approx 1.285714} \end{cases} \qquad \textbf{\textit{pseudorozwiązanie.}}$$

Można też policzyć maksymalny błąd tego wyniku: $B_{\max} = B(x_0, y_0) = 0.285714$

Czasami stosuje się też tzw. *ważoną metodę najmniejszych kwadratów*. Aby zwiększyć lub zmniejszyć wpływ jednego z równań na wynik końcowy, można przypisać każdemu z równań wagę (funkcję lub liczbę) – im większą tym bliżej tej prostej będzie leżało pseudorozwiązanie.
Wprowadza się diagonalną macierz wagową: $W = diag\{w_{ii}\},\ i = 1,2,...,n$ zbierającą wagi przypisane wszystkim równaniom. Odpowiednie modyfikacje ostatecznych układów równań są następujące:

w zapisie wskaźnikowym: $\sum_{i=1}^{n} a_{ik} \sum_{j=1}^{m} w_{ii} a_{ij} x_j = \sum_{i=1}^{n} w_{ii} a_{ik} b_i, \quad k = 1,2,...,m$

w zapisie macierzowym: $\boldsymbol{A^T W A}\, x = \boldsymbol{A^T W}\, b$

Przykład 12
Rozwiązań nadokreślony układ równań z przykładu 11, przypisując każdemu z równań wagę będącą jego numerem kolejnym.

Wagi: $w_{11} = 1, \quad w_{22} = 2, \quad w_{33} = 3$
Funkcjonał błędu: $B(x,y) = 1\cdot(x+y-2)^2 + 2\cdot(x-y)^2 + 3\cdot(x-2y+2)^2$

$$\begin{cases} \dfrac{\partial B}{\partial x} = 2\cdot 1\cdot(x+y-2) + 2\cdot 2\cdot(x-y) + 2\cdot 3\cdot(x-2y+2) = 0 \\ \dfrac{\partial B}{\partial y} = 2\cdot 1\cdot(x+y-2) - 2\cdot 2\cdot(x-y) - 4\cdot 3\cdot(x-2y+2) = 0 \end{cases}$$

$$\begin{cases} 12x-14y=-8 \\ -14x+30y=28 \end{cases} \Rightarrow \begin{cases} \boldsymbol{x_0 = 0.926829} \\ \boldsymbol{y_0 = 1.365854} \end{cases}, \quad B_{\max} = 0.585366$$

WARTOŚCI WŁASNE I WEKTORY WŁASNE MACIERZY

Wartościami własnymi macierzy ***A*** stopnia *n* nazywamy takie wartości $\lambda_1, \lambda_2, ..., \lambda_n$ parametru λ, dla których układ równań

$$\boldsymbol{A}x = \lambda x \tag{1}$$

ma niezerowe rozwiązanie.
Wektor x_r, spełniający przy $\lambda = \lambda_r$ układ równań (1), nazywamy *wektorem własnym macierzy* ***A***. Układ (1) ma niezerowe rozwiązanie wtedy, gdy jego wyznacznik jest równy zero, tzn.

$$(\boldsymbol{A} - \lambda \boldsymbol{I}) = 0$$

Po rozwinięciu powyższego wyznacznika otrzymamy równanie algebraiczne stopnia *n* :

$$a_0 + a_1\lambda + a_2\lambda^2 + ... + (-1)^n \lambda^n = 0$$

zwane *równaniem charakterystycznym macierzy* ***A***. Pierwiastki tego równania są oczywiście wartościami własnymi macierzy ***A***

Przykład 1
Niech

$$\boldsymbol{A} = \begin{bmatrix} 1 & 0 & 0 & 4 \\ 0 & 3 & 2 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 2 \end{bmatrix}$$

Znajdziemy teraz równanie charakterystyczne macierzy ***A***

$$\left|\boldsymbol{A} - \lambda \boldsymbol{I}\right| = \begin{vmatrix} 1-\lambda & 0 & 0 & 4 \\ 0 & 3-\lambda & 2 & 0 \\ 1 & 0 & -\lambda & 0 \\ 1 & 1 & 0 & 2-\lambda \end{vmatrix}$$

Rozwijając ten wyznacznik według elementów pierwszego wiersza, otrzymujemy

$$(1-\lambda)\begin{vmatrix} 3-\lambda & 2 & 0 \\ 0 & -\lambda & 0 \\ 1 & 0 & 2-\lambda \end{vmatrix} - 4\begin{vmatrix} 0 & 3-\lambda & 2 \\ 1 & 0 & \lambda \\ 1 & 1 & 0 \end{vmatrix} =$$

$$= (1-\lambda)(3-\lambda)(-\lambda)(2-\lambda) - 4[(3-\lambda)(-\lambda) + 2] = (\lambda-4)(\lambda-2)(\lambda-1)(\lambda+1)$$

Wartości własne macierzy $\boldsymbol{A}$ są równe $\lambda_1 = 4, \quad \lambda_2 = 2, \quad \lambda_3 = 1, \quad \lambda_4 = -1$.
Aby otrzymać wektory własne, należy rozwiązać układ równań $\boldsymbol{A}x = \lambda x$, gdzie zamiast λ będziemy podstawiać kolejne obliczone wartości własne.
Podstawiając $\lambda = 4$, oraz oznaczając współrzędne wektora własnego przez v_1, v_2, v_3, v_4, otrzymujemy następujący układ

$$\begin{bmatrix} 1 & 0 & 0 & 4 \\ 0 & 3 & 2 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 2 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \end{bmatrix} = 4 \begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \end{bmatrix}$$

lub po rozpisaniu

$$\begin{cases} v_1 + 4v_4 = 4v_1 \\ 3v_2 + 2v_3 = 4v_2 \\ v_1 = 4v_3 \\ v_1 + v_2 + 2v_4 = 4v_4 \end{cases}$$

skąd obliczamy $v_1 = 4v_3, \quad v_2 = 2v_3, \quad v_4 = 3v_3$.
Oczywiście wektor własny nie jest określony jednoznacznie. Jeżeli dodatkowo dokonać jego normalizacji, np. zażądać, aby jego największa współrzędna była równa jedności to wtedy otrzymamy

$$x_1 = (1, \frac{1}{2}, \frac{1}{4}, \frac{3}{4})$$

Podobnie otrzymamy pozostałe wektory własne

$$x_2 = (1, -1, \frac{1}{2}, \frac{1}{4}), \quad x_3 = (1, -1, 1, 0), \quad x_4 = (-1, -\frac{1}{2}, 1, \frac{1}{2})$$

Można oczywiście inaczej znormalizować dany wektor *x*, np. tak, aby jego długość była równa jedności, tzn.

$$\|x\| = \sqrt{v_1^2 + v_2^2 + ... + v_n^2} = 1$$

Podana w powyższym przykładzie metoda znajdywania wartości własnych oraz wektorów własnych jest bardzo pracochłonna, szczególnie w przypadku macierzy wysokiego stopnia.
Dlatego też rzadko rozwiązuje się problem własny macierzy na podstawie definicji. Szczególnie kłopotliwe może być wyznaczenie samych wartości własnych, gdy wielomian występujący w równaniu charakterystycznym nie ma pierwiastków wymiernych.

Przykład 2
Niech

$$\boldsymbol{A}=\begin{bmatrix} 1 & 3 & -1 \\ 1 & 2 & 4 \\ -1 & 2 & 3 \end{bmatrix}.$$

Równanie charakterystyczne

$$(\boldsymbol{A}-\lambda\boldsymbol{I})=\begin{vmatrix} 1-\lambda & 3 & -1 \\ 1 & 2-\lambda & 4 \\ -1 & 2 & 3-\lambda \end{vmatrix}$$

po rozwinięciu (np. względem pierwszego wiersza) ma postać następującego wielomianu

$$\lambda^3-6\lambda^2-\lambda+27=0$$

Wielomian ten nie posiada pierwiastków wymiernych, (co łatwo sprawdzić, gdyż mogłyby one wynosić odpowiednio $\lambda_i=1,3,9,27$ ale żadna z tych liczb nie spełnia równania). Równanie trzeciego stopnia posiada odpowiednie wzory na swoje pierwiastki rzeczywiste, (jeżeli istnieją) – tzw. *wzory Cardana*, ale są one dość uciążliwe w użyciu. Dlatego posłużymy się w tym przypadku *metodami numerycznymi* dla określenia jednego z pierwiastków, aby pozostałe dwa wyznaczyć już w sposób analityczny. Budując z powyższego wielomianu schemat iteracyjny dla *metody Newtona*

$$F(\lambda)=\lambda^3-6\lambda^2-\lambda+27$$
$$\lambda_{n+1}=\lambda_n-\frac{F(\lambda_n)}{F'(\lambda_n)}=\lambda_n-\frac{\lambda_n^3-6\lambda_n^2-\lambda_n+27}{3\lambda_n^2-12\lambda_n-1}$$

oraz startując np. z $\lambda_0=1$ otrzymujemy dla czterech kolejnych iteracji

$$\lambda_1=3.1\quad \lambda_2=2.676414\quad \lambda_3=2.720801\quad \lambda_4=2.721158$$

Ostatni wynik można uznać już za satysfakcjonujący gdyż odpowiadające mu tempo zbieżności $\frac{|\lambda_3-\lambda_4|}{|\lambda_4|}=0.000131$ jest relatywnie małą liczba.

Zatem przyjmujemy do dalszych obliczeń $\lambda=\lambda_4=2.721158$. W celu wyznaczenia pozostałych pierwiastków równania dzielimy wyjściowy wielomian przez $(\lambda-2.721158)$ otrzymując w rezultacie

$$\lambda^3-6\lambda^2-\lambda+27=(\lambda-2.721158)(\lambda^2-3.278842\lambda-9.922247)$$

Równanie kwadratowe rozwiązujemy w znany analityczny sposób wyznaczając pozostałe dwa pierwiastki. Ostatecznie wartości własne macierzy $\boldsymbol{A}$ wynoszą (w kolejności rosnącej)

$$\lambda_1=-1.911628,\quad \lambda_2=2.721158,\quad \lambda_3=5.190470$$

Dla porównania analitycznie policzone wartości własne wynoszą $-1.911629,\ 2.721159,\ 5.190470$. Zatem powyższe wielkości numeryczne są bardzo dobrym przybliżeniem ścisłych wyników analitycznych.

Dalej postępujemy podobnie jak w przykładzie 1 w celu wyznaczenia wektorów własnych. Dla $\lambda_1 = -1.911628$ i odpowiadającego jej wektora $v_1 = (x_1, y_1, z_1)$ budujemy układ równań

$$\begin{bmatrix} 1-\lambda_1 & 3 & -1 \\ 1 & 2-\lambda_1 & 4 \\ -1 & 2 & 3-\lambda_1 \end{bmatrix} \begin{bmatrix} x_1 \\ y_1 \\ z_1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \Rightarrow \begin{bmatrix} 2.911628 & 3 & -1 \\ 1 & 3.911628 & 4 \\ -1 & 2 & 4.911628 \end{bmatrix} \begin{bmatrix} x_1 \\ y_1 \\ z_1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

Do dwóch pierwszych równań (trzecie to tożsamość w stosunku do nich) dołączamy warunek na jednostkową długość wektora własnego.

$$\begin{cases} 2.911628\, x_1 + 3\, y_1 - z_1 = 0 \\ x_1 + 3.911628\, y_1 + 4\, z_1 = 0 \\ x_1^2 + y_1^2 + z_1^2 = 1 \end{cases}$$

Rozwiązanie tego układu daje współrzędne wektora v_1

$$x_1 = 0.723635 \quad y_1 = -0.575143 \quad z_1 = 0.381527$$

Analogiczne obliczenia można przeprowadzić dla pozostałych wartości własnych. Odpowiadające im wektory własne wynoszą

$$v_2 = (x_2, y_2, z_2) \quad x_2 = -0.878231 \quad y_2 = -0.458209 \quad z_2 = 0.136948$$
$$v_3 = (x_3, y_3, z_3) \quad x_3 = 0.423079 \quad y_3 = 0.756967 \quad z_3 = 0.498000$$

W ogólności dla dowolnej macierzy może okazać się, iż dana macierz nie posiada wartości własnych rzeczywistych lub posiada wartości własne wielokrotne. W drugim przypadku nie istnieje jeden unormowany wektor własny, ale cały ich zbiór leżący na konkretnej płaszczyźnie.
Bardzo często występującymi macierzami w naukach technicznych są macierze symetryczne, np. w mechanice ciała odkształcalnego takimi macierzami są macierz naprężeń i macierz odkształceń dla materiału izotropowego. Można wykazać następujące twierdzenie:

Twierdzenie 1
Każda macierz symetryczna dodatnio określona posiada wszystkie wartości własne rzeczywiste dodatnie i różne od siebie.

Przykład 3
Macierz naprężeń dla płaskiego stanu naprężenia opisana jest w każdym punkcie ciała

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

Znaleźć postać macierzy w układzie własnym oraz jej kierunki główne.

Układamy równanie charakterystyczne (tu również nazywane *równaniem wiekowym* lub *sekularnym*)

$$(\boldsymbol{A}-\lambda\boldsymbol{I})=\begin{vmatrix}3-\lambda & \sqrt{2}\\ \sqrt{2} & 2-\lambda\end{vmatrix}=0 \quad\Rightarrow\quad \lambda^2-5\lambda+4=0$$

Wartości własne wynoszą $\lambda_1=1,\quad \lambda_2=4$.
Wektory własne:

- dla $\lambda_1=1$

$$\begin{bmatrix}3-1 & \sqrt{2}\\ \sqrt{2} & 2-1\end{bmatrix}\begin{bmatrix}x_1\\ y_1\end{bmatrix}=\begin{bmatrix}0\\ 0\end{bmatrix} \quad\Rightarrow\quad \begin{cases}2x_1+\sqrt{2}y_1=0\\ x_1^2+y_1^2=1\end{cases}$$

  stąd $x_1=0.816497,\quad y_1=0.577350$.

- dla $\lambda_2=4$

$$\begin{bmatrix}3-4 & \sqrt{2}\\ \sqrt{2} & 2-4\end{bmatrix}\begin{bmatrix}x_2\\ y_2\end{bmatrix}=\begin{bmatrix}0\\ 0\end{bmatrix} \quad\Rightarrow\quad \begin{cases}-x_2+\sqrt{2}y_2=0\\ x_2^2+y_2^2=1\end{cases}$$

  stąd $x_2=-0.577350,\quad y_2=0.816497,.$

Dla wektorów własnych (tu: wersorów wyznaczających osie główne) macierzy symetrycznych istnieje warunek ich wzajemnej ortogonalności

$$\begin{bmatrix}x_1\\ y_1\end{bmatrix}^T\cdot\begin{bmatrix}x_2\\ y_2\end{bmatrix}=\begin{bmatrix}0.816497\\ 0.577350\end{bmatrix}^T\cdot\begin{bmatrix}-0.577350\\ 0.816497\end{bmatrix}=0$$

co sprowadza się do obliczenia iloczynu skalarnego wektorów (dla wektorów prostopadłych iloczyn skalarny jest równy zeru).

W analitycznej i numerycznej analizie problemów własnych macierzy pomocnicze są następujące twierdzenia:

Twierdzenie 2
Jeżeli macierz posiada różne wartości własne to istnieje zbiór liniowo niezależnych wektorów własnych, z dokładnością do stałej, co oznacza istnienie jednoznacznych kierunków tych wektorów.

Twierdzenie 3 *(Cayley – Hamiltona)*
Macierz symetryczna drugiej walencji ($A=\left[a_{ij}\right]$) spełnia swoje własne równanie charakterystyczne.

$$\boldsymbol{A}^3 - I_1^A \boldsymbol{A}^2 + I_2^A \boldsymbol{A} - I_3^A \boldsymbol{I},$$

gdzie I_1^A, I_2^A, I_3^A *są jej niezmiennikami.*

Twierdzenie 4
Jeżeli $g(x)$ *jest wielomianem, a* λ *jest wartością własną macierzy* $\boldsymbol{A}$*, to* $g(\lambda)$ *jest wartością własną macierzy* $g(\boldsymbol{A})$.

Przykład 4
Wartości własne macierzy $\boldsymbol{A}$ wynoszą $\lambda_i = \{-2, 0, 1, 3\}$. Obliczyć wartości własne macierzy $\boldsymbol{B} = \boldsymbol{A}^3 - 2\boldsymbol{A}^2 + \boldsymbol{A} - 10\boldsymbol{I}$

Konsekwencją twierdzenia 4 jest przeniesienie zależności miedzy macierzami $\boldsymbol{A}$ i $\boldsymbol{B}$ na zależność między ich wartościami własnymi, czyli:

$$\lambda_B = \lambda_A^3 - 2\lambda_A^2 + \lambda_A - 10$$

co pozwala bardzo łatwo obliczyć wartości własne macierzy $\boldsymbol{B}$

$$\lambda_1 = (-2)^3 - 2(-2)^2 + (-2) - 10 = -28$$
$$\lambda_2 = 0^3 - 2 \cdot 0^2 + 0 - 10 = -10$$
$$\lambda_3 = 1^3 - 2 \cdot 1^2 + 1 - 10 = -10$$
$$\lambda_4 = 3^3 - 2 \cdot 3^2 + 3 - 10 = 2$$

Twierdzenie 5
Transformacja macierzy $\boldsymbol{A}$ *przez podobieństwo nie zmienia jej wartości własnych.*
Jeżeli $\boldsymbol{R}$ *jest macierzą nieosobliwą to transformacją przez podobieństwo nazywamy przekształcenie* $\boldsymbol{R}^{-1}\boldsymbol{A}\boldsymbol{R}$. *Wartości własne tej nowej macierzy są takie same jak wartości własne macierzy wyjściowej* $\boldsymbol{A}$.

Twierdzenie 6
Transformacja ortogonalna macierzy $\boldsymbol{A}$ *nie zmienia ani jej wartości własnych ani jej ewentualnej symetrii.*
Jeżeli $\boldsymbol{Q}$ *jest macierzą nieosobliwą i taką, że* $\boldsymbol{Q}^T\boldsymbol{Q} = \boldsymbol{I}$ *to transformacją ortogonalna nazywamy przekształcenie* $\boldsymbol{Q}^T\boldsymbol{A}\boldsymbol{Q}$. *Wartości własne tej nowej macierzy są takie same jak wartości własne macierzy wyjściowej* $\boldsymbol{A}$.

Twierdzenie 7 (*Gerszgorina*)
Niech $\boldsymbol{A}$ *będzie macierzą kwadratową o wymiarze n i wyrazach* a_{ij} $(i, j = 1, 2, ..., n)$. *Jeżeli określimy dyski* $u_i, i = 1, 2, ..., n$ *o środkach odpowiadającym wyrazom* a_{ii} *na przekątnej*

głównej macierzy i promieniach R_i, *gdzie* $R_i = \sum_{\substack{k=1 \\ k \neq i}}^{n} |a_{ik}|$ *to widmo macierzy* $\boldsymbol{A}$ *(zbiór wartości własnych) można oszacować poprzez wzory:*

$$\lambda \in \langle \lambda_{\min}, \lambda_{\max} \rangle$$
$$\lambda_{\min} > \min_i (a_{ii} - R_i)$$
$$\lambda_{\max} < \max_i (a_{ii} + R_i)$$

Oszacowania powyższe stają się rzeczywistymi wartościami $\lambda_{\min}$ i $\lambda_{\max}$ dla macierzy ściśle dominującej na przekątnej głównej.
Macierz nazywamy *macierzą ściśle dominującą na przekątnej głównej*, jeżeli:
$|a_{ii}| > \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}|$, $i = 1, 2, ..., n$.

Przykład 5
Oszacować widmo wartości własnych korzystając z *twierdzenia Gerszgorina* dla macierzy:

$$\boldsymbol{A} = \begin{bmatrix} -2 & 1 & 3 \\ -1 & 4 & 2 \\ 3 & -2 & 3 \end{bmatrix}$$

Wyrazy na przekątnej głównej: $a_{11} = -2, \quad a_{22} = 4, \quad a_{33} = 3$.
Promienie dysków: $R_1 = 1 + 3 = 4, \quad R_2 = |-1| + 2 = 3, \quad R_3 = 3 + |-2| = 5$.
Oszacowanie wartości własnych:

$$\lambda_{\min} > \min \begin{bmatrix} -2-4 \\ 4-3 \\ 3-5 \end{bmatrix} = \min \begin{bmatrix} -6 \\ 1 \\ -2 \end{bmatrix} = -6, \quad \lambda_{\min} > -6$$

$$\lambda_{\max} < \max \begin{bmatrix} -2+4 \\ 4+3 \\ 3+5 \end{bmatrix} = \max \begin{bmatrix} 2 \\ 7 \\ 8 \end{bmatrix} = 8, \quad \lambda_{\max} < 8$$

czyli $\lambda \in \langle -6, 8 \rangle$.
W rzeczywistości macierz $\boldsymbol{A}$ ma jedną wartość własną rzeczywistą równą: -2.980286 mieszczącą się w powyższym przedziale.

Jednym z zastosowań powyższego twierdzenia jest jego wykorzystanie do zbadania dodatniej określoności danej macierzy kwadratowej $\boldsymbol{A}$.
Macierz $\boldsymbol{A}$ o wymiarze n nazywamy *macierzą dodatnio określoną*, jeśli jest nieosobliwa ($\det(\boldsymbol{A}) \neq 0$) oraz dla dowolnego wektora $x \in \Re^n$ spełniona jest nierówność $x^T \boldsymbol{A}\, x > 0$.

Ponieważ badanie dodatniej określoności macierzy z definicji jest kłopotliwe, stosuje się to tego różne twierdzenia, oprócz *twierdzenia 1-szego* także:

Twierdzenie 8
*Jeżeli macierz kwadratowa **A** o wyrazach rzeczywistych jest ściśle dominująca na przekątnej głównej i ma dodatnie wyrazy na przekątnej głównej to **A** jest dodatnio określona.*

Często również wykorzystuje się do badania dodatniej określoności macierzy pojęcie podwyznacznika macierzy: jeśli znaki podwyznaczników macierzy (od rzędu *1-szego* aż do rzędu *n-tego*) tworzą naprzemienny ciąg lub są takie same to macierz jest dodatnio określona.

Według *twierdzenia 1-szego*, aby wykazać, że macierz jest dodatnio określona, należy udowodnić, iż jej wartości własne są dodatnie i różne od siebie. Ponieważ *twierdzenie Gerszgorina* oszacowuje widmo macierzy, można go zastosować w celu zbadania pierwszej tezy. Natomiast zbadanie, czy wartości własne są od siebie różne, wymaga zastosowania tzw. *ciągów Sturma* i nie będzie rozważane w tym opracowaniu.

Przykład 6
Wykorzystać *twierdzenie Gerszgorina* do zbadania dodatniej określoności następujących macierzy:

$$\boldsymbol{A}=\begin{bmatrix}2 & 1 & 1\\ 1 & 2 & 1\\ 1 & 1 & 2\end{bmatrix} \qquad \boldsymbol{B}=\begin{bmatrix}3 & -2 & 1\\ -2 & 3 & 2\\ 1 & 2 & 3\end{bmatrix}$$

Dla macierzy ***A***:

$$\lambda_{\min} > \min\begin{bmatrix}2-2\\ 2-2\\ 2-2\end{bmatrix} = \min\begin{bmatrix}0\\ 0\\ 0\end{bmatrix} = 0, \quad \lambda_{\min} > 0$$

$$\lambda_{\max} < \max\begin{bmatrix}2+2\\ 2+2\\ 2+2\end{bmatrix} = \max\begin{bmatrix}4\\ 4\\ 4\end{bmatrix} = 4, \quad \lambda_{\max} < 4$$

$$\Rightarrow \quad \lambda \in (0,4)$$

Wniosek: macierz ***A*** może być dodatnio określona.

Dla macierzy ***B***:

$$\lambda_{\min} > \min\begin{bmatrix}3-3\\ 3-4\\ 3-4\end{bmatrix} = \min\begin{bmatrix}0\\ -1\\ -1\end{bmatrix} = -1, \quad \lambda_{\min} > -1$$

$$\lambda_{\max} < \max\begin{bmatrix}3+3\\ 3+4\\ 3+4\end{bmatrix} = \max\begin{bmatrix}6\\ 7\\ 7\end{bmatrix} = 7, \quad \lambda_{\max} < 7$$

$$\Rightarrow \quad \lambda \in (-1,7)$$

Wniosek: macierz ***B*** nie jest dodatnio określona.

Metody numeryczne do znajdywania wartości i wektorów własnych można podzielić na :

- metody obliczania wszystkich wartości i wektorów własnych (np. *metoda Jacobiego*),
- metody obliczania wartości własnych i odpowiadających im wektorów własnych w z góry określonych pasmach widma wartości własnych,
- metody obliczania pojedynczej wartości własnej i odpowiadającego jej wektora własnego.

W opracowaniu zostaną przedstawione jedynie metody z ostatniej grupy. Większość z nich to metody iteracyjne.

Metoda potęgowa

Jedną z najprostszych metod jednoczesnego obliczania wartości własnych oraz wektorów własnych macierzy $\boldsymbol{A}$ jest następująca metoda iteracyjna.
Przypuśćmy, że wartości własne $\lambda_1, \lambda_2, ..., \lambda_n$ są rzeczywiste i spełniają nierówności $|\lambda_1| > |\lambda_2| > ... > |\lambda_n|$. Wybiera się dowolny wektor y_0, a następnie za pomocą wzoru iteracyjnego $y_{n+1} = A\, y_n$ buduje się ciąg wektorów $y_1, y_2, ...$ Okazuje się, że dla dostatecznie dużych *n*, wektor y_n jest bliski wektorowi własnemu macierzy $\boldsymbol{A}$, odpowiadającemu największej, co do modułu wartości własnej. Wartość własną otrzymamy dzieląc dowolną współrzędną wektora y_{n+1} przez tą samą współrzędną wektora y_n .

Przykład 7
Niech macierz $\boldsymbol{A}$ będzie postaci:

$$\boldsymbol{A} = \begin{bmatrix} 2 & -1 & 0 & 0 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 2 \end{bmatrix}$$

Przyjmijmy wektor początkowy $y_0 = (1, -1, 1, -1)$. Kolejne iteracje dają następujący ciąg wektorów:

$$y_1 = \begin{bmatrix} 2 & -1 & 0 & 0 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 2 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ -1 \\ 1 \\ -1 \end{bmatrix} = \begin{bmatrix} 3 \\ -4 \\ 4 \\ -3 \end{bmatrix} \qquad y_2 = \begin{bmatrix} 2 & -1 & 0 & 0 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 2 \end{bmatrix} \cdot \begin{bmatrix} 3 \\ -4 \\ 4 \\ -3 \end{bmatrix} = \begin{bmatrix} 10 \\ -15 \\ 15 \\ -10 \end{bmatrix} \quad itd.$$

$y_3 = (35, -55, 55, -35)$ $y_4 = (125, -200, 200, -125)$ $y_5 = (450, -725, 725, -405)$
$y_6 = (1625, -2625, 2625, -1625)$ $y_7 = (5875, -9500, 9500, -5875)$
$y_8 = (21250, -34375, 34375, -21250)$
Stosunki odpowiednich współrzędnych wektorów y_8 i y_7 są równe:

$$\frac{5875}{21250}=3.61702,\quad \frac{-9500}{-34375}=3.61842,\quad \frac{9500}{34375}=3.61842,\quad \frac{-5875}{-21250}=3.61702$$

Widzimy, że wszystkie cztery liczby są dość bliskie sobie, stąd wnioskujemy, że każda z nich jest bliska największej, co do modułu wartości własnej macierzy $\boldsymbol{A}$.
Bardziej dokładną wartość własną otrzymamy, jeżeli podzielimy skalarny kwadrat wektora y_8 przez iloczyn skalarny $y_7 \cdot y_8$. Otrzymamy wówczas

$$\lambda_1^* = \frac{y_8 \cdot y_8}{y_7 \cdot y_8} = \frac{21250\cdot 21250+(-34375)\cdot(-34375)+34375\cdot 34375+(-21250)\cdot(-21250)}{5875\cdot 21250+(-9500)\cdot(-34375)+9500\cdot 34375+(-5875)\cdot(-21250)} =$$
$$= 3.61804$$

Odpowiedni wektor własny jest równy $x_1^* = (0.61818, -1, 1, -0.61818)$. Wektor x_1^* jest tak znormalizowany, że jego największa współrzędna jest równa jedności. Gdyby przyjąć kryterium jednostkowej długości wektora własnego, to wynosiłby on wtedy:
$x_1^* = (0.37118, -0.60146, 0.60146, -0.37181)$.
Dokładna wartość największej, co do modułu wartości własnej jest równa $\lambda_1 = 3.618034$.

Metoda Rayleigha

Jest to najpopularniejsza metoda wśród metod iteracyjnych znajdywania wartości własnej macierzy $\boldsymbol{A}$ o wymiarze *n*, największej, co do modułu. Wywodzi się ona z omawianej wyżej *metody potęgowej*, wykorzystuje m.in. własności *twierdzenia 6*, oraz wyrażenie postaci:

$\boldsymbol{A}x = \lambda x \qquad \lambda = \dfrac{x^T \boldsymbol{A} x}{x^T x}$, zwane w literaturze *ilorazem Rayleigha*.

Algorytm metody wygląda następująco:
Poszukiwana jest wartość własna λ największa, co do modułu oraz odpowiadający jej wektor własny *x* (lub unormowany *v*): $\boldsymbol{A}x = \lambda x$
Przyjmujemy na starcie wektor x_0. Przypisujemy $x_{k=0} = x_0$, gdzie *k* oznacza *k-tą* iterację.

- Normalizujemy wektor x_k (dzielimy go przez jego długość):

  $$v_k = \frac{x_k}{\|x_k\|} = \frac{x_k}{\sqrt{x_k^T x_k}}$$

- Obliczamy kolejne przybliżenie wektora własnego $x_{k+1} = \boldsymbol{A} \cdot v_k$.
- Obliczamy *iloraz Rayleigha*:

  $\lambda_{k+1} = \dfrac{v_k^T \boldsymbol{A} v_k}{v_k^T v_k} = v_k^T x_{k+1}$ (jest to po prostu iloczyn skalarny dwóch wektorów będący liczbą – kolejnym przybliżeniem wartości własnej λ_1.

- Obliczamy poziom błędów (począwszy od drugiej iteracji – dla *k = 1*):

  $$\varepsilon_{k+1}^{\lambda} = \left|\frac{\lambda_{k+1}-\lambda_k}{\lambda_{k+1}}\right|$$, norma błędu przy obliczaniu wartości własnej

  $\varepsilon_{k+1}^{v} = \|v_{k+1} - v_k\|$, norma błędu przy obliczaniu wektora własnego.

- Sprawdzamy kryterium przerwania iteracji:

$\varepsilon_{k+1}^{\lambda} \le B_1, \quad \varepsilon_{k+1}^{v} \le B_2$, gdzie B_1, B_2 - zadane poziomy dokładności obydwu wielkości.

Jeżeli powyższe kryterium jest spełnione to: $\lambda_1 \approx \lambda_{k+1}, \quad v_1 \approx v_{k+1}$

Przykład 8
Dana jest macierz:

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

Znaleźć jej wartość własną największą, co do modułu i odpowiadający jej wektor własny korzystając z *metody Rayleigha.*

Wartości własne macierzy wynoszą: $\lambda_1 = 4, \quad \lambda_2 = \lambda_3 = 1$
Przyjmujemy wektor startowy $x_0 = (1,0,0)$

Pierwsza iteracja $k = 0$:

$$\|x_0\| = 1 \quad \Rightarrow \quad v_0 = \frac{x_0}{1} = (1,0,0)$$

$$x_1 = \boldsymbol{A}v_0 = \begin{bmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \\ 1 \end{bmatrix}$$

$$\lambda_1 = v_0^T x_1 = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 2 \\ 1 \\ 1 \end{bmatrix} = 2$$

Druga iteracja $k = 1$:

$$\|x_1\| = 2.499490 \quad \Rightarrow \quad v_1 = \frac{x_1}{2.499490} = (0.816497, 0.408248, 0.408248)$$

$$x_2 = \boldsymbol{A}v_1 = \begin{bmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{bmatrix} \cdot \begin{bmatrix} 0.816497 \\ 0.408248 \\ 0.408248 \end{bmatrix} = \begin{bmatrix} 2.449490 \\ 2.041241 \\ 2.041241 \end{bmatrix}$$

$$\lambda_2 = v_1^T x_2 = \begin{bmatrix} 0.816497, 0.408248, 0.408248 \end{bmatrix} \begin{bmatrix} 2.449490 \\ 2.041241 \\ 2.041241 \end{bmatrix} = 3.666667$$

$$\|x_2\| = 3.785939 \quad \Rightarrow \quad v_2 = \frac{x_2}{3.785939} = (0.649997 \quad 0.539164 \quad 0.539164)$$

$$\varepsilon_2^{\lambda} = \left| \frac{\lambda_2 - \lambda_1}{\lambda_2} \right| = \left| \frac{3.666667 - 2}{2} \right| = 0.454545,$$

$$\varepsilon_2^{v} = \|v_2 - v_1\| = \left\| \begin{bmatrix} 0.649997 \\ 0.539164 \\ 0.539164 \end{bmatrix} - \begin{bmatrix} 0.816497 \\ 0.408248 \\ 0.408248 \end{bmatrix} \right\| = 0.251014$$

Trzecia iteracja $k = 2$:

$$x_3 = \boldsymbol{A}v_2 = \begin{bmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{bmatrix} \cdot \begin{bmatrix} 0.649997 \\ 0.539164 \\ 0.539164 \end{bmatrix} = \begin{bmatrix} 2.372321 \\ 2.264488 \\ 2.264488 \end{bmatrix}$$

$$\lambda_3 = v_2^T x_3 = \begin{bmatrix} 0.649997, 0.539164, 0.539164 \end{bmatrix} \begin{bmatrix} 2.372321 \\ 2.264488 \\ 2.264488 \end{bmatrix} = 3.976744$$

$$\|x_3\| = 3.985439 \quad \Rightarrow \quad v_3 = \frac{x_3}{3.985439} = (0.595247 \quad 0.568190 \quad 0.568190)$$

$$\varepsilon_3^{\lambda} = \left| \frac{\lambda_3 - \lambda_2}{\lambda_3} \right| = \left| \frac{3.976744 - 3.666667}{3.666667} \right| = 0.07797,$$

$$\varepsilon_3^{v} = \|v_3 - v_2\| = \left\| \begin{bmatrix} 0.595247 \\ 0.568190 \\ 0.568190 \end{bmatrix} - \begin{bmatrix} 0.649997 \\ 0.539164 \\ 0.539164 \end{bmatrix} \right\| = 0.066054$$

Już po trzech iteracjach widoczne jest, na jakim poziomie stabilizują się wyniki. Wartość własną z precyzją do sześciu miejsc otrzymano po $k = 7$ iteracjach:

$$\lambda \approx \lambda_7 = 4.0, \quad \varepsilon_7^{\lambda} = 0.000001$$

$$v \approx v_7 = (0.577421, 0.577315, 0.577315), \quad \varepsilon_7^{v} = 0.000259$$

Zaobserwować można szybszą zbieżność samej wartości własnej niż wektora własnego.

Zarówno w przypadku *metody potęgowej* jak i *metody Rayleigha* pozostałe wartości i wektory własne można znaleźć stosując różne modyfikacje tych metod jak np. *metodę iteracji odwrotnej* zbieżną do wartości własnej najbliższej zeru czy *przesunięcie widma macierzy* o zadaną wartość. Stosuje się też zabiegi mające na celu przyśpieszenie zbieżności metod iteracyjnych.

UKŁADY RÓWNAŃ ŹLE UWARUNKOWANYCH

Przy rozwiązywaniu układów równań liniowych postaci $\boldsymbol{A}x = b$ można mieć do czynienia z przypadkiem, gdy

- ***det(A) = 0*** - osobliwość macierzy współczynników powoduje brak rozwiązań przy dowolnym niezerowym wektorze wyrazów wolnych *b* lub tożsamość dla zerowego wektora *b*,
- ***det(A)*** $\neq$ ***0*** - zapewnia istnienie jednoznacznego rozwiązania postaci $x = \boldsymbol{A}^{-1} b$
- ***det(A)*** $\approx$ ***0*** - *układ źle uwarunkowany*. W takiej sytuacji bardzo małe zmiany w wyrazach macierzy współczynników mogą spowodować ogromne zmiany w rozwiązaniu.

W celu zbadania stopnia uwarunkowania układu równań oblicza się tzw. *wskaźnik uwarunkowania k* – liczbę o takiej własności, że

- $k = 1$ - idealne uwarunkowanie,
- $k = \infty$ - układ osobliwy.

Sposoby obliczania *wskaźnika uwarunkowania k* dla macierzy współczynników $\boldsymbol{A}$:

- $k_A = \|\boldsymbol{A}\| \cdot \|\boldsymbol{A}^{-1}\|, \quad \|\boldsymbol{A}\| = \sqrt{\sum_{i=1}^{n} a_{ij}^2}$
- $k_A = \left| \frac{\lambda_{\max}}{\lambda_{\min}} \right|$.

Posługując się ostatnim wzorem można obliczać wartości własne analitycznie (wtedy wzór ma słuszność gł. dla macierzy symetrycznych) lub numerycznie (np. z *twierdzenia Gerszgorina* dla macierzy ściśle dominujących na przekątnej głównej)

Przykład 9

Wykazać, która z macierzy $\boldsymbol{H} = \begin{bmatrix} -1 & -\frac{1}{3} \\ 1 & 1 \end{bmatrix}$ oraz $\boldsymbol{J} = \begin{bmatrix} -1 & -4 \\ -1 & -3 \end{bmatrix}$ jest lepiej uwarunkowana.

Wynik uzasadnić liczbowo.

W zadaniu należy obliczyć osobno wskaźniki uwarunkowania dla każdej z macierzy i sprawdzić, który z nich jest bliższy jedności. Posłużymy się przy obliczaniu wskaźnika kryterium normowym.

Macierze $\boldsymbol{H}^{-1}$ oraz $\boldsymbol{J}^{-1}$ można obliczyć analitycznie (ze *wzoru Gaussa*) lub stosując odpowiedni algorytm numeryczny (*eliminacja Gaussa*, rozkład na czynniki trójkątne, metody iteracyjne). Ponieważ wymiary macierzy są małe, ich odwrotności obliczono analitycznie.

$$\det(\boldsymbol{H}) = -\frac{2}{3} \quad \Rightarrow \quad \boldsymbol{H}^{-1} = -\frac{3}{2}\begin{bmatrix} 1 & \frac{1}{3} \\ -1 & -1 \end{bmatrix}$$

$$\det(\boldsymbol{J}) = -1 \quad \Rightarrow \quad \boldsymbol{J}^{-1} = -\begin{bmatrix} -3 & 4 \\ 1 & -1 \end{bmatrix}$$

Odpowiednie normy średniokwadratowe wynoszą:

$$\|\boldsymbol{H}\| = \sqrt{1+\frac{1}{9}+1+1} = \sqrt{\frac{28}{9}} = \frac{2}{3}\sqrt{7}, \qquad \|\boldsymbol{H}^{-1}\| = \frac{3}{2}\sqrt{1+\frac{1}{9}+1+1} = \frac{3}{2}\sqrt{\frac{28}{9}} = \sqrt{7}$$

$$\|\boldsymbol{J}\| = \sqrt{1+16+1+9} = \sqrt{27} = 3\sqrt{3}, \qquad \|\boldsymbol{J}^{-1}\| = \sqrt{9+16+1+1} = 3\sqrt{3}$$

zaś wskaźniki uwarunkowania :

$$k_H = \|\boldsymbol{H}\| \cdot \|\boldsymbol{H}^{-1}\| = \frac{2}{3}\sqrt{7} \cdot \sqrt{7} = \frac{14}{3} \approx\sim 4.666667$$

$$k_J = \|\boldsymbol{J}\| \cdot \|\boldsymbol{J}^{-1}\| = 3\sqrt{3} \cdot 3\sqrt{3} = 27$$

Ponieważ $|k_H - 1| < |k_J - 1|$ to lepiej uwarunkowana jest macierz $\boldsymbol{H}$.

Kryterium związanego ze ścisłym wyznaczeniem wartości własnych nie można zastosować, gdyż macierz $\boldsymbol{J}$ nie posiada rzeczywistych rozwiązań problemu własnego. Oszacowanie widm macierzy z twierdzenia Gerszgorina daje w rezultacie:

- dla macierzy $\boldsymbol{H}$:

$$\lambda_{\min} \approx \min(\begin{bmatrix} -1 \\ 1 \end{bmatrix} - \begin{bmatrix} \frac{1}{3} \\ 1 \end{bmatrix}) = -\frac{4}{3}, \qquad \lambda_{\max} \approx \max(\begin{bmatrix} -1 \\ 1 \end{bmatrix} + \begin{bmatrix} \frac{1}{3} \\ 1 \end{bmatrix}) = 2$$

$$k_H = \left|\frac{2}{\frac{-4}{3}}\right| = \frac{3}{2} = 1.5$$

- dla macierzy $\boldsymbol{J}$:

$$\lambda_{\min} \approx \min(\begin{bmatrix} -1 \\ -3 \end{bmatrix} - \begin{bmatrix} 4 \\ 1 \end{bmatrix}) = -5, \qquad \lambda_{\max} \approx \max(\begin{bmatrix} -1 \\ -3 \end{bmatrix} + \begin{bmatrix} 4 \\ 1 \end{bmatrix}) = 3$$

$$k_J = \left|\frac{3}{-5}\right| = \frac{3}{5} = 0.6$$

Oszacowanie okazało się fałszywe (macierze nie są ściśle dominujące na przekątnej głównej).

Przykład 10

Zbadać uwarunkowanie macierzy

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

Macierz spełnia wymagania kryteria do stosowania wzoru opartego na wartościach własnych. Równanie charakterystyczne wynosi: $\lambda^2 - 4\lambda + 3$ a wartości własne: $\lambda_{\max} = 3, \quad \lambda_{\min} = 1$

Wskaźnik uwarunkowania: $k_A = \left| \frac{\lambda_{\max}}{\lambda_{\min}} \right| = 3$.

Na podstawie wskaźnika można ustalić, z jaką precyzją należy podać elementy macierzy $\boldsymbol{A}$ aby uzyskać żądaną dokładność rozwiązania. Służy do tego wzór :

$q \approx p - \log(k)$,

gdzie : q – liczba cyfr znaczących elementów macierzy, p – dokładność rozwiązania.

Np. dla $p = 6$ mamy $p = q + \log(k) = 6 + \log(3) = 6.47 \approx 7$ miejsc znaczących współczynników macierzy $\boldsymbol{A}$.

Uwarunkowanie macierzy można poprawić stosując większą precyzję obliczeń lub tzw. *metody regularyzacji*.

Opracował: **mgr Sławomir Milewski**
Samodzielny Zakład Metod Komputerowych w Mechanice L6, WIL, PK

I. APROKSYMACJA I INTERPOLACJA FUNKCJI JEDNEJ ZMIENNEJ

Ogólnie zagadnienie aproksymacji można opisać następująco:

- Dane są punkty należące bądź to do wykresu funkcji bądź pochodzące z danych eksperymentalnych lub numerycznych (liczba punktów – n)

  $(x_i, f_i) \quad dla \quad i = 1, 2, ..., n$

  Odcięte x_i nazywamy węzłami aproksymacji, natomiast rzędne f_i wartościami węzłowymi.
- Przyjmuje się tzw. rząd aproksymacji m $(m = 0, 1, ..., n-1)$. Jest to ilość niezależnych liniowo funkcji bazowych $\varphi_i(x)$, przyjmowanych na podstawie danego kryterium, a także ilość nieznanych współczynników liczbowych a_i, które zostaną wyznaczone w dalszym ciągu zadania. Ogólny zapis funkcji aproksymującej:

$$p(x) = a_0\varphi_0(x) + a_1\varphi_1(x) + ... + a_m\varphi_m(x) = \sum_{i=0}^{m} a_i\varphi_i(x) \tag{1}$$

  lub w notacji macierzowej:

$$p(x) = \boldsymbol{a}^T\boldsymbol{\varphi}(x), \quad gdzie: \underset{(1\times m)}{\boldsymbol{a}} = \begin{bmatrix} a_0 \\ a_1 \\ ... \\ a_m \end{bmatrix}, \quad \underset{(1\times m)}{\boldsymbol{\varphi}}(x) = \begin{bmatrix} \varphi_0(x) \\ \varphi_1(x) \\ ... \\ \varphi_m(x) \end{bmatrix}$$

- Przyjmuje się tzw. wagi w_i dla każdego węzła z osobna, które świadczą o odejściu krzywej aproksymacyjnej od oryginalnej wartości węzłowej wg zależności: im większa waga, tym bliżej tego właśnie punktu przejdzie krzywa. Wagi można dobierać np. według kryterium odległościowego od ustalonego z góry punktu. Wagi zbiera się do macierzy diagonalnej zwanej macierzą wagową.

$$\underset{(n\times n)}{\boldsymbol{W}} = diag(w_i) = \begin{bmatrix} w_1 & 0 & 0 & 0 \\ 0 & w_2 & 0 & 0 \\ 0 & 0 & ... & 0 \\ 0 & 0 & 0 & w_n \end{bmatrix}.$$

  Oczywiście wprowadzanie wag nie jest konieczne. W takim przypadku: $w_1 = w_2 = ... = w_n = 1$.
- Wyznacza się współczynniki liczbowe a_i z następującego układu równań:

$$\underset{(n\times m)}{\boldsymbol{\Phi}} = \left|\varphi_j(x_i)\right| = \begin{bmatrix} \varphi_0(x_1) & \varphi_1(x_1) & \ldots & \varphi_m(x_1) \\ \varphi_0(x_2) & \varphi_1(x_2) & \ldots & \varphi_m(x_2) \\ \ldots & \ldots & \ldots & \ldots \\ \varphi_0(x_n) & \varphi_1(x_n) & \ldots & \varphi_m(x_n) \end{bmatrix}, \qquad \underset{(1\times n)}{\boldsymbol{F}} = \left|f_i\right| = \begin{bmatrix} f_1 \\ f_2 \\ \ldots \\ f_n \end{bmatrix}$$

$\boldsymbol{\Phi}^{\mathbf{T}}\boldsymbol{W}\boldsymbol{\Phi}\boldsymbol{a} = \boldsymbol{\Phi}^{\mathbf{T}}\boldsymbol{W}\ \boldsymbol{\Phi}, \quad \Rightarrow \quad \boldsymbol{a} = (\boldsymbol{\Phi}^{\mathbf{T}}\boldsymbol{W}\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^{\mathbf{T}}\boldsymbol{W}\ \mathbf{F}$

Na ich podstawie można budować aproksymację funkcji za pomocą wzoru (1).

- Oblicza się błąd aproksymacji na podstawie następujących wzorów:
  - Dla aproksymacji ciągłej: $\varepsilon = \int_{x_1}^{x_n} (p(x) - f(x))dx$,
  - Dla aproksymacji dyskretnej:

$$\varepsilon = \left\| p(x_i) - f_i \right\|_{i=1,2,\ldots,n} = \begin{cases} \sqrt{\sum_{i=1}^{n} (p(x_i) - f_i)^2}, & \text{dla normy Euklidesa} \\ \max_i \left| p(x_i) - f_i \right|, & \text{dla normy maksimowej} \end{cases}.$$

Powyższy algorytm aproksymacji jest ogólny i prawdziwy dla dowolnej liczby węzłów, ilości i postaci funkcji bazowych. Wszystkie poniższe rodzaje aproksymacji można łatwo wyprowadzić korzystając z tego algorytmu. Jest on jednak dość uciążliwy zwłaszcza w obliczeniach ręcznych, stąd dla konkretnego rodzaju aproksymacji korzysta się z innych zależności, prostszych w zapisie i zastosowaniu.

INTERPOLACJA FUNKCJI

Interpolacja funkcji to taka aproksymacja, w której funkcja $p(x)$ przechodzi przez wszystkie punkty (x_i, f_i), $i = 1, 2, \ldots, n$ bez żadnego wyjątku. To znaczy, iż błąd liczony jak dla aproksymacji dyskretnej musi być w węzłach bezwarunkowo równy zeru. Stąd warunek interpolacji formułuje się następująco:

$p(x_i) = f_i$, dla $i = 1, 2, \ldots, n$.

Implikuje to od razu postać funkcji interpolacyjnej:

$$p(x) = \sum_{i=1}^{n} a_i \varphi_i(x) \tag{2}$$

, tzn., że funkcji bazowych (oraz współczynników interpolacji) musi być dokładnie tyle, ile węzłów. Tak, więc zadanie interpolacji jest zadaniem jednoznacznym (jest tylko jedna krzywa interpolacyjna, która dla danego zestawu funkcji bazowych przechodzi ściśle przez wszystkie dane punkty). W zapisie macierzowym interpolacja wygląda następująco:

$$p(x) = \boldsymbol{a}^T \boldsymbol{\varphi}(x), \quad gdzie: \ \underset{(1\times n)}{\boldsymbol{a}} = \begin{bmatrix} a_1 \\ a_2 \\ \ldots \\ a_n \end{bmatrix}, \quad \underset{(1\times n)}{\boldsymbol{\varphi}}(x) = \begin{bmatrix} \varphi_1(x) \\ \varphi_2(x) \\ \ldots \\ \varphi_n(x) \end{bmatrix}.$$

Współczynniki a_i wyznacza się z następującego układu równań:

$$\underset{(n\times n)}{\mathbf{\Phi}} = \left|\varphi_j(x_i)\right| = \begin{bmatrix} \varphi_1(x_1) & \varphi_2(x_1) & \ldots & \varphi_n(x_1) \\ \varphi_1(x_2) & \varphi_2(x_2) & \ldots & \varphi_n(x_2) \\ \ldots & \ldots & \ldots & \ldots \\ \varphi_1(x_n) & \varphi_2(x_n) & \ldots & \varphi_n(x_n) \end{bmatrix}, \qquad \underset{(1\times n)}{\mathbf{F}} = \left|f_i\right| = \begin{bmatrix} f_1 \\ f_2 \\ \ldots \\ f_n \end{bmatrix}$$

$$\mathbf{\Phi}\,\boldsymbol{a} = \mathbf{F}, \quad \Rightarrow \quad \boldsymbol{a} = \mathbf{\Phi}^{-1}\,\mathbf{F}$$

Powyższy układ równań ma jedno rozwiązanie, gdy macierz $\mathbf{\Phi}$ jest nieosobliwa, a to zachodzi wtedy, gdy węzły interpolacji nie pokrywają się (wyjściowe przyporządkowanie dyskretne jest funkcją).

W przypadku interpolacji zwężanie po liczbie funkcji bazowych nie jest konieczne, gdyż jest ona równa liczbie węzłów, a więc macierz współczynników $\mathbf{\Phi}$ jest od samego początku macierzą kwadratową. Nie ma sensu również wprowadzać wag, gdyż z założenia wynika, iż w węzłach krzywa ma mieć ustalone z góry wartości, a więc sterowanie jej przebiegiem w węzłach jest niemożliwe (wprowadzanie wag nie będzie miało żadnego wpływu na wynik końcowy). Po wyznaczeniu współczynników można budować krzywą wg wzoru (2).
Błąd interpolacji zależy od wyboru funkcji bazowych.

Należy również nadmienić, iż interpolacja podana w tej postaci nie jest najlepszą z możliwych interpolacji, mimo iż przechodzi przez wszystkie dane punkty. Kosztem tego jest jej niestabilne i niczym nie kontrolowane zachowanie między węzłami. Interpolacja słabo więc odtwarza oryginalną funkcję. Im więcej węzłów, tym większych niestabilności można się spodziewać, zwłaszcza dla interpolacji wielomianowej. Poza tym, przejście funkcji przez wszystkie punkty ściśle wcale nie musi być najlepszym rozwiązaniem, zwłaszcza przy obróbce danych eksperymentalnych, gdy każdy wynik obarczony jest błędem zupełnie zaniedbywanym w wyniku zastosowania interpolacji.

1. Interpolacja jednomianowa

Jest to najprostsza, ale i najbardziej prymitywna z interpolacji (wymaga rozwiązywania dużych układów równań). Znana jest w klasycznej postaci: dane jest kilka punktów, przez które ma przejść krzywa. Zapisuje się więc jej wzór wielomianowy zależny od tylu współczynników, ile jest punktów, przez które ma ona przejść. Współczynniki znajduje się z układu równań, powstałego z zapisania jej przejścia ścisłego przez wszystkie punkty. Np. dla dwóch punktów $(x_1, f_1), (x_2, f_2)$ zapisuje się wzór funkcji liniowej $p(x) = a\,x + b$, a współczynniki a i b znajduje się z warunków $p(x_1) = f_1$ oraz $p(x_2) = f_2$. Dokładnie to samo postępowanie wyniknie z ogólnego schematu interpolacji, tylko ze szczególną postacią funkcji bazowych w postaci kolejnych jednomianów:

$$\varphi_1(x) = 1, \quad \varphi_2(x) = x, \quad \varphi_3(x) = x^2, \quad \varphi_4(x) = x^3, \quad \ldots, \quad \varphi_n(x) = x^{n-1}.$$

Ogólnie: $\varphi_i(x) = x^{i-1}, \quad dla \;\; i = 1, 2, \ldots, n$.

Krzywą (2) znajduje się wtedy z układu równań:

$$\mathbf{\Phi a} = \mathbf{F}, \quad \Rightarrow \quad \boldsymbol{a} = \mathbf{\Phi}^{-1}\,\mathbf{F}, \text{ gdzie: } \underset{(n\times n)}{\mathbf{\Phi}} = \begin{bmatrix} 1 & x_1 & \dots & x_1^{n-1} \\ 1 & x_2 & \dots & x_2^{n-1} \\ 1 & \dots & \dots & \dots \\ 1 & x_n & \dots & x_n^{n-1} \end{bmatrix}$$

Macierz $\mathbf{\Phi}$ przy interpolacji jednomianowej w literaturze nosi nazwę macierzy Van Der Monda. Podobnie jak przy ogólnym sformułowaniu interpolacji, macierz $\mathbf{\Phi}$ jest nieosobliwa $(\det \mathbf{\Phi} \neq 0)$, gdy $\forall_{i,j}\ x_i \neq x_j$.

Przykład 1
Dany jest zbiór punktów:

i	1	2	3
x_i	0	1	2
f_i	0	1	4

Dokonać interpolacji jednomianowej.

Dobieramy trzy funkcje bazowe: $\varphi_1(x) = 1, \quad \varphi_2(x) = x, \quad \varphi_3(x) = x^2$. Przyjmujemy postać interpolacji $p(x) = \sum_{i=1}^{3} a_i \varphi_i(x) = a_1\varphi_1(x) + a_2\varphi_2(x) + a_3\varphi_3(x) = a_1 + a_2 x + a_3 x^2$.

Budujemy macierz Van Der Monda: $\mathbf{\Phi} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 2 & 4 \end{bmatrix}$ oraz układ równań:

$$\begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 2 & 4 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 4 \end{bmatrix} \quad \Rightarrow \quad \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}.$$

Stąd: $p(x) = a_1 + a_2 x + a_3 x^2 = 0 + 0 \cdot x + 1 \cdot x^2 = x^2$.
Interpolacja idealnie odtworzyła pierwotną parabolę, z której zdjęte zostały punkty.

2. Interpolacja Lagrange'a

W przypadku, gdy funkcjami bazowymi są wielomiany coraz wyższych stopni, wynik końcowy (krzywa interpolacyjna) jest oczywiście taki sam. Natomiast można poszukiwać go na różne sposoby. Jeden z nich pozwala na ominięcie rozwiązywania układu równań zakładając specyficzną wielomianową postać funkcji bazowych. Otóż, jeżeli przyjmie się funkcje bazowe ($\varphi_i(x) \equiv L_i(x)$, tzw. *wielomiany Lagrange'a*) w zależności od rozłożenia węzłów tak, że: $L_i(x_j) = \begin{cases} 0, & i \neq j \\ 1, & i = j \end{cases}$, to macierz współczynników $\mathbf{\Phi}$ przyjmie następującą postać:

$$\mathbf{\Phi} = \begin{bmatrix} L_1(x_1) & L_2(x_1) & \dots & L_n(x_1) \\ L_1(x_2) & L_2(x_2) & \dots & L_n(x_2) \\ \dots & \dots & \dots & \dots \\ L_1(x_n) & L_2(x_n) & \dots & L_n(x_n) \end{bmatrix} = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 \end{bmatrix} = \mathbf{I}$$

Układ równań będzie miał rozwiązanie: $\mathbf{I\,a = F} \quad \Rightarrow \quad \mathbf{a = F}$.

Tak więc w przypadku tej interpolacji (tzw. *interpolacji Lagrange'a*) przy odpowiednim doborze funkcji bazowych znane są od razu współczynniki krzywej interpolacyjnej – są nimi wartości węzłowe:

$$p(x) = \sum_{i=1}^{n} f_i\, L_i(x) = f_1\, L_1(x) + f_2\, L_2(x) + \dots + f_n\, L_n(x)\,.$$

Jedyną trudność stanowi więc znalezienie wielomianów Lagrange'a. Jest ich tyle, ile węzłów. Dowolny, *i-ty* wielomian zeruje się we wszystkich węzłach oprócz węzła z numerem *i-tym*, w którym przyjmuje wartość *1*. Oczywiście pomiędzy węzłami wielomian przyjmuje wartości niezerowe. Można go opisać wzorem (tzw. *wzór interpolacyjny Lagrange'a*):

$$L_i(x) = \frac{(x-x_1)\cdot(x-x_2)\cdot\dots\cdot(x-x_{i-1})\cdot(x-x_{i+1})\cdot\dots\cdot(x-x_n)}{(x_i-x_1)\cdot(x_i-x_2)\cdot\dots\cdot(x_i-x_{i-1})\cdot(x_i-x_{i+1})\cdot\dots\cdot(x_i-x_n)} = \frac{\prod_{\substack{j=1 \\ j\neq i}}^{n}(x-x_j)}{\prod_{\substack{j=1 \\ j\neq i}}^{n}(x_i-x_j)}\,.$$

Licznik jest iloczynem różnic $(x-x_j)$ tworzonym z pominięciem węzła x_i. Pojawia się on za to w mianowniku, który jest licznikiem policzonym dla $x = x_i$.

Błąd interpolacji Lagrange'a dla dowolnego *x* można określić z następującego wzoru:

$$\varepsilon(x) = \frac{f^{(n)}(\xi)\cdot\prod_{i=1}^{n}(x-x_i)}{n!} \le \left|\frac{f^{(n)}_{\max}\cdot\prod_{i=1}^{n}(x-x_i)}{n!}\right|, \quad x_1 \le \xi \le x_n\,.$$

$f^{(n)}$ oznacza pochodną *n-tego* rzędu, natomiast ξ jest punktem pośrednim z przedziału, w którym dokonuje się interpolacji.

Uogólnieniem interpolacji Lagrange'a jest *interpolacja l'Hermitte'a*, w której w węzłach obok wartości funkcji mogą być również dane wartości pochodnych.

Przykład 2

Dany jest zbiór punktów:

i	1	2	3
x_i	0	1	2
f_i	0	1	4

Dokonać interpolacji Lagrange'a.

Budujemy kolejne wielomiany Lagrange'a:

$$L_1(x)=\frac{(x-x_2)(x-x_3)}{(x_1-x_2)(x_1-x_3)}=\frac{(x-1)(x-2)}{(0-1)(0-2)}=\frac{1}{2}(x-1)(x-2)$$
$$L_2(x)=\frac{(x-x_1)(x-x_3)}{(x_2-x_1)(x_2-x_3)}=\frac{(x-0)(x-2)}{(1-0)(1-2)}=-x(x-2)$$
$$L_3(x)=\frac{(x-x_1)(x-x_2)}{(x_3-x_1)(x_3-x_2)}=\frac{(x-0)(x-1)}{(2-0)(2-1)}=\frac{1}{2}x(x-1)$$

Wzór interpolacyjny:

$$p(x)=f_1L_1(x)+f_2L_2(x)+f_3L_3(x)$$
$$p(x)=0\cdot\frac{1}{2}(x-1)(x-2)+1\cdot(-x)(x-2)+4\cdot\frac{1}{2}x(x-1)=-x^2+2x+2x^2-2x=x^2$$

Błąd interpolacji jest równy *0* dla dowolnego *x* z uwagi, iż pochodna rzędu *n = 3* wyjściowej funkcji $f(x)=x^2$ jest równa $f^{III}(x)\equiv 0$.

Przykład 3

Dokonać interpolacji Lagrange'a funkcji ciągłej $f(x)=\sin(x)$ w przedziale $\langle 2,4\rangle$ stosując różne liczby węzłów. Wyznaczyć błąd interpolacji. Obliczyć wartość wielomianu interpolacyjnego dla $x_0=\pi$ dla i porównać z wynikiem ścisłym.

W podanym przedziale dokonujemy dyskretyzacji funkcji za pomocą $n=3$ węzłów równomiernie rozłożonych. Otrzymujemy następujące punkty:

i	1	2	3
x_i	2	3	4
$f_i=\sin(x_i)$	0.909297	0.141120	-0.756802

Budujemy wielomiany Lagrange'a:

$$L_1(x)=\frac{(x-3)(x-4)}{(2-3)(2-4)}=\frac{1}{2}(x-3)(x-4),\quad L_2(x)=\frac{(x-2)(x-4)}{(3-2)(3-4)}=-(x-2)(x-4),$$
$$L_3(x)=\frac{(x-2)(x-3)}{(4-2)(4-3)}=\frac{1}{2}(x-2)(x-3)$$

Budujemy interpolację:

$$p(x)=0.909297\cdot\frac{1}{2}(x-3)(x-4)+0.141120\cdot(-1)\cdot(x-2)(x-4)-0.756802\cdot\frac{1}{2}\cdot(x-2)(x-3)=$$
$$=-0.06487x^2-0.443815x+2.056416$$

Wartość interpolacji dla $x_0=\pi$: $p_0=p(x_0=\pi)=0.021828$.

Wartość ścisła dla $x_0=\pi$: $f_0=0.0$.

Błąd bezwzględny wyniku: $\varepsilon_0=|p_0-f_0|=|0-0.021828|=0.021828$.

Oszacowanie błędu interpolacji:

$$f^{III}(x)=-\sin(x),\quad f^{III}_{\max}=f^{III}(x=2)=-0.909297$$

$$\varepsilon(x) \le \left| -0.909297 \frac{(x-2)(x-3)(x-4)}{6} \right| = \left| 0.151550 \cdot (x-2)(x-3)(x-4) \right|$$

Błąd interpolacji dla $x_0 = \pi$ wynosi:

$\varepsilon_0 = \varepsilon(x_0 = \pi) \le \left| 0.151550 \cdot (\pi - 2)(\pi - 3)(\pi - 4) \right| = 0.02103$.

Wynik ulega istotnej poprawie dla większej liczby węzłów: dla $n = 4$ $p_0 = 0.000404$,a dla $n = 5$ $p_0 = 0.000256$.

3. Odwrotna interpolacja Lagrange'a

Zamiast budować interpolację na zmiennych niezależnych *x*, można odwrócić miejscami zmienne *x* z *y* i znaleźć w rezultacie wielomian interpolacyjny *p(y)*:

Dla danych punktów węzłowych: (x_i, f_i), $i = 1, 2, ..., n$ budujemy *n* wielomianów Lagrange'a, ale traktując *y* jako zmienną niezależną:

$$L_i(y) = \frac{(y-f_1)\cdot(y-f_2)\cdot\ldots\cdot(y-f_{i-1})\cdot(y-f_{i+1})\cdot\ldots\cdot(y-f_n)}{(f_i-f_1)\cdot(f_i-f_2)\cdot\ldots\cdot(f_i-f_{i-1})\cdot(f_i-f_{i+1})\cdot\ldots\cdot(f_i-f_n)} = \frac{\prod_{\substack{j=1 \\ j\neq i}}^{n}(y-f_j)}{\prod_{\substack{j=1 \\ j\neq i}}^{n}(f_i-f_j)}$$

oraz stosujemy zmodyfikowany wzór interpolacyjny Lagrange'a:

$$p(y) = \sum_{i=1}^{n} x_i L_i(y) = x_1 L_1(y) + x_2 L_2(y) + \ldots + x_n L_n(y)$$

Teraz można odtworzyć, jaki oryginalnie x_0 był przypisany danemu y_0 poprzez obliczenie $x_0 = p(y_0)$. Metoda odwrotna może być też dobrym przybliżeniem metod iteracyjnych do znajdywania pierwiastka równania algebraicznego $f(x) = 0$. Wtedy budując interpolację odwrotną na zbiorze punktów funkcji $f(x)$ w przedziale $a \le x \le b$ można oszacować z dobrym przybliżeniem miejsce zerowe oryginalnej funkcji $f(x)$ poprzez obliczenie $x^* = p(y = 0)$.

Uwaga! Warunkiem rozwiązywalności zadania jest różnowartościowość funkcji *f(x)*.

Przykład 4

Znaleźć przybliżenie miejsca zerowego równania $x - \sin(x) = 0$ w przedziale $x \in (\frac{\pi}{2}, \frac{3}{2}\pi)$.

W podanym przedziale wprowadzamy $n = 3$ węzły, dobierając wartości węzłowe na podstawie równania $f(x) = x - \sin(x)$.

i	1	2	3
x_i	$\frac{\pi}{2}$	π	$\frac{3}{2}\pi$
$f_i = f(x_i)$	$\frac{\pi}{2} - 1$	π	$\frac{3}{2}\pi + 1$

Budujemy na wartościach węzłowych wielomiany Lagrange'a:

$$L_1(y)=\frac{(y-f_2)(y-f_3)}{(f_1-f_2)(f_1-f_3)}=\frac{(y-\pi)(y-\frac{3}{2}\pi-1)}{(\frac{\pi}{2}-1-\pi)(\frac{\pi}{2}-1-\frac{3}{2}\pi-1)}=\frac{2}{(\pi+2)^2}(y-\pi)(y-\frac{3}{2}\pi-1)$$

$$L_2(y)=\frac{(y-f_1)(y-f_3)}{(f_2-f_1)(f_2-f_3)}=\frac{(y-\frac{\pi}{2}+1)(y-\frac{3}{2}\pi-1)}{(\pi-\frac{\pi}{2}+1)(\pi-\frac{3}{2}\pi-1)}=-\frac{4}{(2+\pi)^2}(y-\frac{\pi}{2}+1)(y-\frac{3}{2}\pi-1)$$

$$L_3(y)=\frac{(y-f_1)(y-f_2)}{(f_3-f_1)(f_3-f_2)}=\frac{(y-\pi)(y-\frac{\pi}{2}+1)}{(\frac{3}{2}\pi+1-\pi)(\frac{3}{2}\pi+1-\frac{\pi}{2}+1)}=\frac{2}{(\pi+2)^2}(y-\pi)(y-\frac{\pi}{2}+1)$$

oraz wzór interpolacyjny:

$$p(y)=x_1L_1(y)+x_2L_2(y)+x_3L_3(y)$$

$$p(y)=\frac{\pi}{2}\cdot\frac{2}{(\pi+2)^2}(y-\pi)(y-\frac{3}{2}\pi-1)+\pi\cdot(-1)\frac{4}{(2+\pi)^2}(y-\frac{\pi}{2}+1)(y-\frac{3}{2}\pi-1)$$

$$+\frac{3}{2}\pi\cdot\frac{2}{(\pi+2)^2}(y-\pi)(y-\frac{\pi}{2}+1)=\pi\frac{y+2}{\pi+2}$$

Przybliżenie miejsca zerowego równania: $x^*\approx p(0)=\frac{2\pi}{2+\pi}=1.222031$.

4. Wielomiany Czebyszewa

Interpolacja wielomianowa funkcji dyskretnej daje wyniki ścisłe, gdy interpolowany jest wielomian, co najwyżej stopnia *n-1*. Dla stopni wyższych oraz dla wyjściowych funkcji niebędących wielomianami wyniki są w jakiś sposób przybliżone. Dla wysokich stopni interpolacji krzywe wielomianowe są niestabilne, tzn. mimo przejścia ścisłego przez wszystkie punkty między nimi zaczynają coraz bardziej się rozbiegać do nieskończoności. Aby zapewnić maksymalną stabilność takich wyników stosuje się jako funkcje bazowe wielomiany ortogonalne (lub ortogonalne z wagą) np. funkcje specjalne Lagrange'a (nie mylić z wcześniej omawianymi wielomianami Lagrange'a), l'Hermitte'a, Legendre'a czy Czebyszewa. Te ostatnie mają jeszcze jedną bardzo ważną dla aproksymacji własność: jeżeli mianowicie tak dobierze się węzły aproksymacji, aby były one równe miejscom zerowym odpowiedniego *wielomianu Czebyszewa*, to wtedy maksymalny błąd tak zbudowanej interpolacji wielomianowej zostanie zminimalizowany:

Błąd maksymalny interpolacji: $\varepsilon(x)\le\left|f^{(n)}_{\max}\right|\cdot\left|\prod_{i=1}^{n}(x-x_i)\right|$

Znaleźć minimum maksymalnej wartości w przedziale $\langle -1,1\rangle$ z iloczynu $\prod_{i=1}^{n}(x-x_i)$, czyli: $\min_{x_i}\max_{-1\le x\le 1}\left|\prod_{i=1}^{n}(x-x_i)\right|$ - oryginalne *zagadnienie Czebyszewa*.

Wielomiany Czebyszewa można określić na dwa sposoby:

- Sposób iteracyjny: $T_n(x) = \cos(n \cdot arc\cos x)$,
- Sposób rekurencyjny: $\begin{cases} T_0(x) = 1 \\ T_1(x) = x \\ T_n(x) = 2 \cdot x \cdot T_{n-1}(x) - T_{n-2}(x) \end{cases}$

Powyższe wzory obowiązują w przedziale $-1 \le x \le 1$. To przedział, w którym wielomiany Czebyszewa są określone i w którym są ortogonalne.
W konkretnych zastosowaniach bardziej korzystny jest wzór rekurencyjny, gdzie dany wielomian oblicza się na podstawie dwóch poprzednich. Dla przykładu pokazano kilka następnych wielomianów Czebyszewa:

$T_2(x) = 2 \cdot x \cdot T_2(x) - T_1(x) = 2 \cdot x \cdot x - 1 = 2x^2 - 1$

$T_3(x) = 2 \cdot x \cdot T_3(x) - T_2(x) = 2 \cdot x \cdot (2x^2 - 1) - x = 4x^3 - 3x$

$T_4(x) = 8x^4 - 8x^2 + 1$

$T_5(x) = 16x^5 - 20x^3 + 5$

Aby znaleźć miejsca zerowe *n-tego* wielomianu Czebyszewa, nie trzeba rozwiązywać w tym celu równania $T_n(x) = 0$; można posłużyć się gotowym wzorem:

$$x_i = \cos\frac{2 \cdot i + 1}{n}\frac{\pi}{2}, \quad i = 0,1,...,n-1.$$

Własność ortogonalności wielomianów Czeybszewa z wagą $\mu(x) = \frac{1}{\sqrt{1-x^2}}$ polega na tym, iż całka:

$$I_{ij} = \int_{-1}^{1} \frac{T_i(x) \cdot T_j(x)}{\sqrt{1-x^2}} dx = \begin{cases} 0, & i \neq j \\ \frac{\pi}{2}, & i = j \neq 0 \\ \pi, & i = j = 0 \end{cases}.$$

Ponieważ w konkretnych zadaniach mamy do czynienia z dowolnym przedziałem *x*, dlatego też zachodzi często potrzeba transformacji wyjściowego przedziału do przedziału, w którym znane są wielomiany Czebyszewa i odwrotnie:
Niech $z \in \langle a,b \rangle, \quad x \in \langle -1,1 \rangle$:

- Przejście $z \to x \quad : \quad x = \frac{2z - (b+a)}{b-a}$,
- Przejście $x \to z \quad : \quad z = \frac{1}{2}[(b-a) \cdot x + (b+a)]$.

Uwaga! W zadaniach interpolacji można bazować na zadanej siatce węzłów a tylko jako funkcji bazowych użyć wielomianów Czebyszewa (tzw. *interpolacja Czebyszewa*), albo przyjąć węzły jako miejsca zerowe odpowiedniego wielomianu Czebyszewa a interpolować używając do tego jednej z poznanych metod (w tym także interpolacji Czebyszewa). To samo dotyczy także aproksymacji funkcji.

Przykład 5

Dana jest funkcja dyskretna (z_i, f_i), $i = 1,2,3$, taka jak w przykładach 1 i 2:

i	1	2	3
z_i	0	1	2
f_i	0	1	4

Dokonać interpolacji Czebyszewa.

Węzłów nie wyznaczamy – są z góry podane. Do interpolacji na trzech węzłach potrzebne będą trzy wielomiany Czebyszewa (w przedziale $x \in \langle -1,1 \rangle$):

$T_0(x) = 1, \quad T_1(x) = x, \quad T_2(x) = 2x^2 - 1.$

Wzory na transformację między przedziałami $z \in \langle 0,2 \rangle, \quad x \in \langle -1,1 \rangle$: $x = z - 1, \quad z = x + 1.$

Wielomiany Czebyszewa w przedziale $z \in \langle 0,2 \rangle$:

$T_0(z) = 1, \quad T_1(z) = z - 1, \quad T_2(z) = 2(z-1)^2 - 1 = 2z^2 - 4z + 1.$

Tworzymy układ równań:

$$\mathbf{\Phi} = \begin{bmatrix} T_0(z_1) & T_1(z_1) & T_2(z_1) \\ T_0(z_2) & T_1(z_2) & T_2(z_2) \\ T_0(z_3) & T_1(z_3) & T_2(z_3) \end{bmatrix} = \begin{bmatrix} 1 & -1 & 1 \\ 1 & 0 & -1 \\ 1 & 1 & 1 \end{bmatrix}, \qquad \mathbf{F} = \begin{bmatrix} f_1 \\ f_2 \\ f_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 4 \end{bmatrix}$$

i rozwiązujemy:

$$\mathbf{\Phi} \cdot \mathbf{a} = \mathbf{F} \quad \Rightarrow \quad \mathbf{a} = \begin{bmatrix} 1.5 \\ 2 \\ 0.5 \end{bmatrix}$$

Wzór interpolacyjny:

$$p(x) = a_0 T_0(z) + a_1 T_1(z) + a_2 T_2(z) = \frac{3}{2} \cdot 1 + 2 \cdot (z-1) + \frac{1}{2}(2z^2 - 4z + 1) = z^2$$

Otrzymany wzór odtwarza pierwotną parabolę, tak samo jak w przypadku interpolacji jednomianowej i Lagrange'a.

Przykład 6

Dokonać interpolacji funkcji $f(z) = \sqrt{1 + z^2}$ w przedziale $z \in \langle 0,5 \rangle$. Jako funkcje bazowe przyjąć wielomiany Czebyszewa, a jako węzły interpolacji miejsca zerowe wielomianu $T_3(x)$.

Zacznijmy od węzłów interpolacji w przedziale $x \in \langle -1,1 \rangle$. Wielomian $T_3(x)$ ma trzy miejsca zerowe, co od razu implikuje trzy węzły a więc interpolację parabolą. Korzystamy ze wzoru na miejsca zerowe:

$$x_i = \cos\frac{2\cdot i+1}{3}\frac{\pi}{2}, \quad i = 0,1,2$$

$$\begin{cases} x_0 = \cos\frac{2\cdot 0+1}{3}\frac{\pi}{2} = \cos\frac{\pi}{6} = \frac{\sqrt{3}}{2} = 0.866025 \\ x_1 = \cos\frac{2\cdot 1+1}{3}\frac{\pi}{2} = \cos\frac{\pi}{2} = 0 \\ x_2 = \cos\frac{2\cdot 0+1}{3}\frac{\pi}{2} = \cos\frac{5}{6}\pi = -\frac{\sqrt{3}}{2} = -0.866025 \end{cases}$$

Natomiast wielomiany potrzebne do wzoru interpolacyjnego:

$$T_0(x) = 1, \quad T_1(x) = x, \quad T_2(x) = 2x^2 - 1$$

Wzory na transformację między przedziałami $z \in \langle 0,5 \rangle$, $x \in \langle -1,1 \rangle$:

$$x(z) = \frac{2}{5}z - 1, \quad z(x) = \frac{5}{2}(x+1).$$

Miejsca zerowe i wielomiany w przedziale $z \in \langle 0,5 \rangle$:

$$z_0 = \frac{5}{2}(x_0+1) = \frac{5}{4}(\sqrt{3}+2) = 4.665064$$

$$z_1 = \frac{5}{2}(x_1+1) = \frac{5}{2} = 2.50$$

$$z_2 = \frac{5}{2}(x_2+1) = \frac{5}{4}(2-\sqrt{3}) = 0.334936$$

$$T_0(z) = 1, \quad T_1(z) = \frac{2}{5}z - 1, \quad T_2(z) = 2\cdot(\frac{2}{5}x-1)^2 - 1 = \frac{8}{25}z^2 - \frac{8}{5}z + 1$$

Dyskretyzacja funkcji $f(z) = \sqrt{1+z^2}$ (węzły ułożono w kolejności rosnącej):

i	1	2	3
z_i	0.334936	2.50	4.665064
$f_i = f(z_i)$	1.054600	2.692582	4.771040

Budowa i rozwiązanie układu równań:

$$\mathbf{\Phi} = \begin{bmatrix} T_0(z_0) & T_1(z_0) & T_2(z_0) \\ T_0(z_1) & T_1(z_1) & T_2(z_1) \\ T_0(z_2) & T_1(z_2) & T_2(z_2) \end{bmatrix} = \begin{bmatrix} 1 & -0.866025 & 0.5 \\ 1 & 0 & -1 \\ 1 & 0.866025 & 0.5 \end{bmatrix}, \qquad \mathbf{F} = \begin{bmatrix} f_1 \\ f_2 \\ f_3 \end{bmatrix} = \begin{bmatrix} 1.054600 \\ 2.692582 \\ 4.771040 \end{bmatrix}$$

$$\mathbf{\Phi}\cdot\mathbf{a} = \mathbf{F} \quad \Rightarrow \quad \mathbf{a} = \begin{bmatrix} 2.839407 \\ 2.145689 \\ 0.146825 \end{bmatrix}.$$

Wzór interpolacyjny:

$$p(z) = a_0\,T_0(z) + a_1\,T_1(z) + a_2\,T_2(z) = 2.839407\cdot 1 + 2.145689\cdot(\frac{2}{5}z - 1) + 0.146825\cdot(\frac{8}{25}z^2 - \frac{8}{5}z + 1)$$

$$= 0.046984\cdot z^2 + 0.623355\cdot z + 0.840544$$

Sprawdzenie własności interpolacyjnych wielomianu $p(z)$:

$p_1 = p(z_1 = 0.334936) = 1.054600 = f_1,$

$p_2 = p(z_2 = 2.50) = 2.692582 = f_{2,}$

$p_3 = p(z_3 = 4.665064) = 4.771040 = f_3$

Obliczenie średniego błędu interpolacji:

$$\varepsilon_{avr} = \int_0^5 [\,p(z) - f(z)]dz = \int_0^5 (0.046984\cdot z^2 + 0.623355\cdot z + 0.840544 - \sqrt{1+z^2})dz = 0.048560$$

Oszacowanie maksymalnego błędu interpolacji:

$$f^{III}(z) = -3\frac{z}{(1+z^2)^{\frac{5}{2}}}, \quad f^{III}_{\max} = f^{III}(z = \frac{1}{2}) = -0.85865$$

$$\varepsilon(z) \le \left| -0.85865\frac{(z-0.334936)(z-2.50)(z-4.665064)}{6} \right| =$$

$$= \left| 0.143108\cdot(z-0.334936)(z-2.50)(z-4.665064) \right|$$

Np. dla $z = 2$ oryginalna wartość funkcji wynosi $f = \sqrt{1+2^2} = 2.236068$, a ta pochodząca z interpolacji $p = p(2) = 2.27519$. Oszacowanie błędu $\varepsilon(2) \le 0.317521$.

5. Interpolacja funkcjami sklejanymi (funkcje typu *spline*)

Przy wzroście liczby węzłów interpolacja daje niepożądane efekty międzywęzłowe w postaci coraz większych gradientów funkcji interpolującej. Aby temu zapobiec i jednocześnie zachować własności interpolacyjne, wprowadzono interpolację funkcjami sklejanymi. Polega ona na znalezieniu krzywej niskiego stopnia, składającej się z różnych kawałków, (czyli o różnych wzorach analitycznych) na przedziałach wyznaczonych przez kolejne pary węzłów. Dodatkowo wymaga się odpowiednich warunków ciągłości: funkcja sklejana (spline) rzędu *k* ma we wszystkich przedziałach wszystkie pochodne ciągłe aż do rzędu *k-1* włącznie.

Rozważmy zbiór punktów $(x_i, f_i), \quad i = 1, 2, ..., n$. Każdy spline rzędu *k* ma pierwszym odcinku $x \in \langle x_1, x_2 \rangle$ wzór: $p(x) = a_k \cdot x^{k-1} + a_{k-1} \cdot x^{k-2} + ... + x \cdot a_2 + a_1 = \sum_{i=1}^{k+1} a_i x^{k+1-i}$. Następnie wraz z przekraczaniem kolejnych węzłów dochodzą następujące składniki wielomianowe:

$p(x) + b_2(x - x_2)^k$ dla $x \in \langle x_2, x_3 \rangle$

$p(x) + b_2(x - x_2)^k + b_3(x - x_3)^k$ dla $x \in \langle x_3, x_4 \rangle$ itd.

Ogólnie spline rzędu *k* można zapisać jednym ogólnym wzorem:

$$s(x) = p(x) + \sum_{i=2}^{n-1} b_i (x - x_i)_+^k = \sum_{i=1}^{k+1} a_i x^{k+1-i} + \sum_{i=2}^{n-1} b_i (x - x_i)_+^k, \quad (x - x_i)_+^k = \begin{cases} (x - x_i)^k, & \text{dla } x > x_i \\ 0, & \text{dla } x \le x_i \end{cases}$$

W każdym spinie są niewiadome współczynniki $a_i,\ \ i=1,2,...,k+1$ i $b_i,\ \ i=2,3,...,n-1$. Razem niewiadomych jest $n-1+k$. Począwszy od $k=2$ (kiedy niewiadomych jest $n-1+2=n+1$) same równania pochodzące od punktów przez które krzywa ma przejść są niewystarczające. Wprowadza się więc dodatkowe warunki na pochodne spline'u w węzłach. I tak spline rzędu *k=1* (spline liniowy) nie wymaga znajomości żadnych dodatkowych warunków), spline rzędu *k=2* (spline kwadratowy, paraboliczny) wymaga znajomości wartości pochodnej w którymś z węzłów, tj. $s^I(x_j)=\alpha$, natomiast spline rzędu *k=3* wymaga znajomości wartości pierwszej i drugiej pochodnej w wybranych dwóch węzłach (może być w tym samym), tj. $s^I(x_j)=\alpha,\ s^{II}(x_l)=\beta$ ($j,l\in\{1,2,...,n\}$). Jeżeli informacje o pochodnych są podane w węzłach pierwszego przedziału $x\in\langle x_1,x_2\rangle$ (tam gdzie obowiązuje przepis $s(x)=p(x)$), to współczynniki a_i można wyznaczyć niezależnie (z układu równań) od współczynników b_i (ze wzoru rekurencyjnego). Jeżeli natomiast warunki brzegowe nie pozwalają na jednoznaczne wyznaczenie odcinka krzywej w przedziale $x\in\langle x_1,x_2\rangle$, to wtedy nie można wyznaczyć rekurencyjnie współczynników b_i, lecz trzeba zbudować w ten sposób układ równań na niewiadome współczynniki a_i i b_i. Dalej rozważany będzie przypadek pierwszy: wszystkie wartości pochodnych dane są w pierwszym węźle ($x=x_1$).

Ogólne wzory na spline (dla $k=1,2,3$):

- Spline liniowy: $s(x)=a_1x+a_2+\sum_{i=2}^{n-1}b_i(x-x_i)_+$,
- Spline kwadratowy: $s(x)=a_1x^2+a_2x+a_3+\sum_{i=2}^{n-1}b_i(x-x_i)_+^2$,
- Spline sześcienny: $s(x)=a_1x^3+a_2x^2+a_3x+a_4+\sum_{i=2}^{n-1}b_i(x-x_i)_+^3$.

Wyznaczenie współczynników $a_i,\ \ i=1,2,...,k+1$:

- Poprzez zapisanie warunków interpolacji spline'u na pierwszym przedziale $x\in\langle x_1,x_2\rangle$ oraz poprzez wykorzystanie ewentualnych dodatkowych informacji o pochodnych w tych węzłach:
  - Dla spline'u liniowego: $\begin{cases}s(x_1)=f_1\\ s(x_2)=f_2\end{cases}\Rightarrow\begin{cases}a_1x_1+a_2=f_1\\ a_1x_2+a_2=f_2\end{cases}\Rightarrow\begin{cases}a_1=...\\ a_2=...\end{cases}$
  - Dla spline'u kwadratowego:
    $$\begin{cases}s(x_1)=f_1\\ s(x_2)=f_2\\ s^I(x_1)=\alpha\end{cases}\Rightarrow\begin{cases}a_1x_1^2+a_2x_1+a_3=f_1\\ a_1x_2^2+a_2x_2+a_3=f_2\\ 2a_1x_1+a_2=\alpha\end{cases}\Rightarrow\begin{cases}a_1=...\\ a_2=...\\ a_3=...\end{cases}$$
  - Dla spline'u sześciennego:
    $$\begin{cases}s(x_1)=f_1\\ s(x_2)=f_2\\ s^I(x_1)=\alpha\\ s^{II}(x_1)=\beta\end{cases}\Rightarrow\begin{cases}a_1x_1^3+a_2x_1^2+a_3x_1+a_4=f_1\\ a_1x_2^3+a_2x_2^2+a_3x_2+a_4=f_2\\ 3a_1x_1^2+2a_2x_1+a_3=\alpha\\ 6a_1x_1+2a_2=\beta\end{cases}\Rightarrow\begin{cases}a_1=...\\ a_2=...\\ a_3=...\\ a_4=...\end{cases}$$

Wyznaczenie współczynników $b_i, \quad i = 2,3,...,n-1$:

- Ze wzoru rekurencyjnego niezależnie od rzędu spline'u; wzór wyprowadza się wykorzystując pozostałe warunki na spline począwszy od $x = x_3$:

  dla $x = x_3$: $s(x_3) = p(x_3) + b_2(x_3 - x_2)^k = f_3 \quad \rightarrow \quad b_2 = \dfrac{f_3 - p(x_3)}{(x_3 - x_2)^k}$

  dla $x = x_4$: $s(x_4) = p(x_4) + b_2(x_4 - x_2)^k + b_3(x_4 - x_3)^k = f_4 \quad \rightarrow \quad b_3 = \dfrac{f_4 - p(x_4) - b_2(x_4 - x_2)^k}{(x_4 - x_3)^k}$

  itd. Ogólnie dla $x = x_{j+1}, \quad j = 2,3,...,n-1$:

  $$s(x_{j+1}) = p(x_{j+1}) + \sum_{i=2}^{j} b_i (x_{j+1} - x_i)^k = f_{j+1} \quad \rightarrow \quad p(x_{j+1}) + \sum_{i=2}^{j-1} b_i (x_{j+1} - x_i)^k + b_j (x_{j+1} - x_j)^k = f_{j+1}$$

  $$b_j = \frac{f_{j+1} - p(x_{j+1}) - \sum_{i=2}^{j-1} b_i (x_{j+1} - x_i)^k}{(x_{j+1} - x_j)^k}.$$

Przykład 7
Dla danych z poprzednich przykładów znaleźć spline liniowy.

i	1	2	3
x_i	0	1	2
f_i	0	1	4

Wzór ogólny spline'u: $s(x) = p(x) + \sum_{i=2}^{3-1} b_i (x - x_i)_+ = a_1 x + a_2 + b_2 (x-1)_+$.

Wyznaczenie współczynników a_1, a_2:

$$\begin{cases} s(0) = 0 \\ s(1) = 1 \end{cases} \Rightarrow \begin{cases} a_2 = 0 \\ a_1 + a_2 = 1 \end{cases} \Rightarrow \begin{cases} a_1 = 1 \\ a_2 = 0 \end{cases} \Rightarrow \quad p(x) = x$$

Wyznaczenie współczynnika b_2:

$s(2) = 4 \quad \Rightarrow \quad p(2) + b_2(2-1) = 4 \quad \Rightarrow \quad b_2 = 4 - 2 = 2$

Wyznaczenie wzoru na spline:

$$s(x) = x + 2 \cdot (x-1)_+ = \begin{cases} x, & dla \;\; 0 \le x \le 1 \\ 3x - 2, & dla \;\; 1 < x \le 2 \end{cases}.$$

Przykład 8
Dla danych z poprzedniego przykładu znaleźć spline kwadratowy.

i	1	2	3
x_i	0	1	2
f_i	0	1	4

Dołączamy informację o pochodnej spline'u dla $x = 0 \rightarrow s^I(0) = \alpha = 0$.

Wzór ogólny spline'u: $\begin{cases} s(x) = p(x) + \sum_{i=2}^{3-1} b_i (x - x_i)_+^2 = a_1 x^2 + a_2 x + a_3 + b_2 (x-1)_+^2 \\ s^I(x) = p^I(x) + 2\sum_{i=2}^{3-1} b_i (x - x_i)_+ = 2a_1 x + a_2 + 2b_2 (x-1)_+ \end{cases}$.

Wyznaczenie współczynników a_1, a_2, a_3:

$$\begin{cases} s(0) = 0 \\ s(1) = 1 \\ s^I(0) = 0 \end{cases} \Rightarrow \begin{cases} a_3 = 0 \\ a_1 + a_2 + a_3 = 1 \\ a_2 = 0 \end{cases} \Rightarrow \begin{cases} a_1 = 1 \\ a_2 = 0 \\ a_3 = 0 \end{cases} \Rightarrow \quad p(x) = x^2$$

Wyznaczenie współczynnika b_2:

$s(2) = 4 \quad \Rightarrow \quad p(2) + b_2 (2-1)^2 = 4 \quad \Rightarrow \quad b_2 = 4 - 2^2 = 0$

Wyznaczenie wzoru na spline:

$s(x) = x^2 + 0 \cdot (x-1)_+^2 = x^2$ dla $0 \le x \le 2$.

W ostatnim przykładzie tylko pozornie interpolacja jest sklejana. Ponieważ dane pochodzą od funkcji kwadratowej, to spline kwadratowy przeistoczył się w oryginalną funkcję o jednym przepisie dla wszystkich *x*.

6. Najlepsza aproksymacja

Aproksymacja to takie dopasowanie krzywej *p(x)* stopnia *m-tego* ($m \le n-1$) do zestawu danych punktów (x_i, f_i), $i = 1, 2, ..., n$, że krzywa aproksymacyjna w ogólności przez żaden punkt ściśle nie przejdzie, dopuszczając odchyłkę między oryginalną wartością f_i, a wartością na krzywej $p(x_i) \neq f_i$. Ogólnym założeniem podejścia *najlepszej aproksymacji* jest minimalizacja sumarycznego błędu (sumy odchyłek) w sensie jakieś normy. Jeżeli zastosowaną normą jest norma Euklidesa (średnio kwadratowa) to metoda nazywa się *metodą najmniejszych kwadratów*.

Aproksymacja: $p(x) = \sum_{i=0}^{m} a_i \varphi_i(x)$.

Błąd aproksymacji: $\varepsilon(x) = f(x) - p(x)$, dla $x_1 \le x \le x_n$.

Najlepsza aproksymacja: $\min_{a_i} \|\varepsilon(x)\| = \min_{a_i} \left\| f(x) - \sum_{i=0}^{m} a_i \varphi_i(x) \right\|$

- Metoda min-max: $\|\varepsilon(x)\|_\infty = \max |\varepsilon(x)| \quad \rightarrow \quad \min_{a_i} \max_x |f(x) - p(x)|$,
- Metoda najmniejszych kwadratów: $\min_{a_i} \|\varepsilon(x)\|_2$:
  - Dla zbioru ciągłego: $\|\varepsilon(x)\|_2 = (\int_{x_1}^{x_n} \varepsilon^2(x) dx)^{\frac{1}{2}}$,
  - Dla zbioru dyskretnego: $\|\varepsilon(x)\|_2 = (\sum_{i=1}^{n} \varepsilon^2(x_i))^{\frac{1}{2}}$.

Najpopularniejszą bazę funkcji bazowej dla aproksymacji stanowią wielomiany, w tym najchętniej używa się funkcji ortogonalnych (lub przynajmniej ortogonalnych wagą), takich

jak wielomiany Czebyszewa, Bessela, Legendre'a czy Hankela. Korzysta się też z bazy jednomianowej, zwłaszcza dla aproksymacji dyskretnej. O jednomianach jako funkcjach bazowych będzie dalej mowa. Funkcja aproksymująca będzie miała wtedy postać:

$$p(x)=\sum_{i=0}^{m} a_i x^{m-i} = a_0 x^m + a_1 x^{m-1} + ... + a_{m-1}x + a_m$$

Współczynniki liczbowe $a_i,\ \ i=0,2,...,m$ należy wyznaczyć na podstawie minimalizacji sumarycznego błędu w każdym z węzłów w sensie normy średnio kwadratowej.

Układamy funkcjonał zbierający informacje o wszystkich węzłach do jednego wzoru:

$$B(a_0,a_1,...,a_m)=(p(x_1)-f_1)^2+(p(x_2)-f_2)^2+...+(p(x_n)-f_n)^2=\sum_{i=1}^{n}(p(x_i)-f_i)^2$$

$$B(a_0,a_1,...,a_m)=\sum_{i=1}^{n}(p(x_i)-f_i)^2=\sum_{i=1}^{n}(\sum_{j=0}^{m}a_j x_i^{m-j}-f_i)^2$$

W celu wyznaczenia niewiadomych współczynników układamy równania będące pochodnymi powyższego funkcjonału względem każdego z nich:

$$\frac{\partial}{\partial a_k}B(a_0,a_1,...,a_m)=2\sum_{i=1}^{n}(\sum_{j=0}^{m}a_j x_i^{m-j}-f_i)x_k^{m-k}=0,\quad \text{dla}\ \ k=0,1,2,...,m$$

Z układu równań $(m+1)\times(m+1)$ wyznaczamy współczynniki, a następnie wyznaczamy *p(x)*:

$$\sum_{i=1}^{n}(\sum_{j=0}^{m}a_j x_i^{m-j})\cdot x_k^{m-k}=\sum_{i=1}^{n}f_i\cdot x_k^{m-k}\quad\Rightarrow\quad\begin{cases}a_0=...\\ a_1=...\\ \quad ...\\ a_m=...\end{cases}\quad\Rightarrow\quad p(x)=\sum_{i=0}^{m}a_i x^{m-i}\ .$$

Zmodyfikowana metoda ważona polega na przypisaniu każdemu z węzłów liczby (wagi) $w_i,\ \ i=1,2,...,n$ świadczącej o stopniu odejścia krzywej od wartości węzłowej: im waga większa waga, tym w rezultacie bliżej krzywa przejdzie obok punktu z tą wagą. Funkcjonał wzbogacony o wagi wygląda następująco:

$$B(a_0,a_1,...,a_m)=w_1\cdot(p(x_1)-f_1)^2+w_2\cdot(p(x_2)-f_2)^2+...+w_n\cdot(p(x_n)-f_n)^2=\sum_{i=1}^{n}w_i\cdot(p(x_i)-f_i)^2$$

Dalsze operacje są identyczne, co prowadzi do układu równań ($k=0,1,2,...,m$):

$$\sum_{i=1}^{n}(w_i\cdot\sum_{j=0}^{m}a_j x_i^{m-j})\cdot x_k^{m-k}=\sum_{i=1}^{n}w_i\cdot f_i\cdot x_k^{m-k}\quad\Rightarrow\quad\begin{cases}a_0=...\\ a_1=...\\ \quad ...\\ a_m=...\end{cases}\quad\Rightarrow\quad p(x)=\sum_{i=0}^{m}a_i x^{m-i}\ .$$

O błędzie aproksymacji decyduje wartość funkcjonału dla policzonych współczynników. Informuje o maksymalnej odchyłce dla danego zestawu węzłów.

Przykład 9
Dla danych z poprzedniego przykładu znaleźć aproksymację linową. Rozpatrzyć dwa przypadki: metodę zwykłą i ważoną przypisując każdemu z węzłów jego numer jako wagę.

i	1	2	3
x_i	0	1	2
f_i	0	1	4

Przyjmujemy funkcję liniową: $p(x) = a \cdot x + b$.

I. Metoda zwykła

Układamy funkcjonał:

$$B(a,b) = \sum_{i=1}^{3} (a \cdot x_i + b - f_i)^2 = (a \cdot 0 + b - 0)^2 + (a \cdot 1 + b - 1)^2 + (a \cdot 2 + b - 4)^2 .$$

Różniczkujemy po zmiennych *a* i *b*:

$$\begin{cases} \dfrac{\partial}{\partial a} B(a,b) = 2 \cdot (a + b - 1) + 2 \cdot 2 \cdot (2a + b - 4) = 0 \\ \dfrac{\partial}{\partial b} B(a,b) = 2 \cdot b + 2 \cdot (a + b - 1) + 2 \cdot (2a + b - 4) = 0 \end{cases}$$

$$\begin{cases} 5a + 3b = 9 \\ 3a + 3b = 5 \end{cases} \Rightarrow \begin{cases} a = 2 \\ b = -\dfrac{1}{3} \end{cases} \Rightarrow \quad p(x) = 2x - \frac{1}{3}.$$

Wyniki zestawiono w tabelce.

i	x_i	f_i	$p_i = p(x_i)$	$\varepsilon_i = p_i - f_i$	ε_i^2
1	0	0	-0.333333	-0.333333	0.111111
2	1	1	1.666667	0.666667	0.444444
3	2	4	3.666667	-0.333333	0.111111

Błąd maksymalny $B_{\max} = \sum_{i=1}^{3} \varepsilon_i^2 = 0.666667$.

II. Metoda ważona

Wagi: $w_1 = 1, \quad w_2 = 2, \quad w_3 = 3$.

$$B(a,b) = \sum_{i=1}^{3} w_i (a \cdot x_i + b - f_i)^2 = 1 \cdot (a \cdot 0 + b - 0)^2 + 2 \cdot (a \cdot 1 + b - 1)^2 + 3 \cdot (a \cdot 2 + b - 4)^2 .$$

Różniczkujemy po zmiennych *a* i *b*:

$$\begin{cases} \dfrac{\partial}{\partial a} B(a,b) = 2 \cdot 2 \cdot (a + b - 1) + 2 \cdot 2 \cdot 3 \cdot (2a + b - 4) = 0 \\ \dfrac{\partial}{\partial b} B(a,b) = 1 \cdot 2 \cdot b + 2 \cdot 2 \cdot (a + b - 1) + 2 \cdot 3 \cdot (2a + b - 4) = 0 \end{cases}$$

$$\begin{cases} 14a + 8b = 26 \\ 8a + 6b = 14 \end{cases} \Rightarrow \begin{cases} a = 2.2 \\ b = -0.6 \end{cases} \Rightarrow \quad p(x) = 2.2x - 0.6 .$$

i	x_i	f_i	$p_i = p(x_i)$	$\varepsilon_i = p_i - f_i$	ε_i^2
1	0	0	-0.60	-0.60	0.360
2	1	1	1.60	0.60	0.360
3	2	4	3.80	0.20	0.04

$$B_{\max} = \sum_{i=1}^{3} w_i \varepsilon_i^2 = 1 \cdot 0.36 + 2 \cdot 0.36 + 3 \cdot 0.04 = 1.20 .$$

Widać poprawę tam gdzie waga była największa: dla węzła $x_3 = 2$.

Przykład 10
Dla danych z poprzedniego zadania zastosować aproksymację kwadratową.

i	1	2	3
x_i	0	1	2
f_i	0	1	4

Funkcja aproksymująca: $p(x) = a \cdot x^2 + b \cdot x + c$.

$$B(a,b,c) = \sum_{i=1}^{3} (a \cdot x_i^2 + b \cdot x_i + c - f_i)^2 = (c-0)^2 + (a+b+c-1)^2 + (4a+2b+c-4)^2 .$$

$$\begin{cases} \dfrac{\partial B}{\partial a} = (a+b+c-1) + 4 \cdot (4a+2b+c-4) = 0 \\ \dfrac{\partial B}{\partial b} = (a+b+c-1) + 2 \cdot (4a+2b+c-4) = 0 \\ \dfrac{\partial B}{\partial c} = c + (a+b+c-1) + (4a+2b+c-4) = 0 \end{cases}$$

$$\begin{cases} 17a+9b+5c=17 \\ 9a+5b+3c=9 \\ 5a+3b+3c=5 \end{cases} \Rightarrow \begin{cases} a=1 \\ b=0 \\ c=0 \end{cases} \Rightarrow \quad p(x) = x^2 .$$

Jest to przypadek szczególny: budowanie aproksymacji kwadratowej na trzech węzłach daje w rezultacie interpolację: otrzymaliśmy wyjściową parabolę. Nie ma sensu stosować metody ważonej.

II. NUMERYCZNE RÓŻNICZKOWANIE FUNKCJI

Wynikiem numerycznego różniczkowania nie jest analityczny wzór na pochodną, ale jej wartość w wybranym węźle zwanym węzłem centralnym. Zadanie sprowadza się do wyznaczenia tzw. *wzoru różnicowego*, czyli wzoru liczącego określoną pochodną w węźle centralnym na podstawie wartości dyskretnych funkcji w innych węzłach, np.:

Dane są wartości funkcji w_0, w_1, w_2 w równych odstępach h. Należy zbudować wzory różnicowe na pierwszą i drugą pochodną w węźle centralnym w_1.

Najbardziej oczywistym sposobem, ale najbardziej prymitywnym jest dokonanie interpolacji (ogólnie: aproksymacji) w podanych punktach, a następnie na podstawie otrzymanego wzoru interpolacyjnego (np. wielomianowego) określić wzór na pochodną i w końcu policzyć

wartość pochodnej w żądanym węźle. Jest to dość złożony proces, gdyż wymaga przejścia z wartości dyskretnych funkcji do wzoru ciągłego a następnie ponowne przejście na wartości dyskretne. Można tego uniknąć, skoro i tak wychodząc od wartości w punktach, szukamy również wartości dyskretnej. Najlepszą metodą do tego celu jest *metoda współczynników nieoznaczonych* bazująca na rozwijaniu wszystkich wartości węzłowych w *szereg Taylora*.

Przyjmujemy lokalny układ współrzędnych w węźle centralnym w_1. Teraz odległości od pozostałych węzłów wynoszą odpowiednio $-h$ oraz h. Rozwijamy każdą z wartości w *szereg Taylora* wokół węzła centralnego zachowując tyle wyrazów ile niewiadomych będzie w końcowym układzie równań. Liczba niewiadomych jest równa ilości informacji, na jakich budujemy wzór różnicowy (w tym przypadku zachowamy trzy wyrazy). Wzoru różnicowego szukamy jako kombinacji liniowej wartości węzłowych i nieznanych (nieoznaczonych – stąd nazwa metody) współczynników liczbowych.

- Dla pierwszej pochodnej: $w^I(x) \approx \sum_{i=0}^{2} a_i w_i = a_0 w_0 + a_1 w_1 + a_2 w_2$,
- Dla drugiej pochodnej: $w^{II}(x) \approx \sum_{i=0}^{2} b_i w_i = b_0 w_0 + b_1 w_1 + b_2 w_2$.

Dla obydwu pochodnych wypisujemy rozwinięcia w poszczególnych węzłach:

$$\begin{cases} w_0 = w_1 - h \cdot w_1 + \frac{1}{2} h^2 w_1 + \ldots \\ w_1 \equiv w_1 \\ w_2 = w_1 + h \cdot w_1 + \frac{1}{2} h^2 w_1 + \ldots \end{cases}$$

Rozwinięcia mnożymy przez współczynniki stojące we wzorach różnicowych. Następnie sumujemy je ze sobą, porządkując wyrazy stojące przy odpowiednich pochodnych. Układ równań powstaje przez porównanie współczynników stojących przy odpowiednich pochodnych: ścisłej pochodnej i wzoru różnicowego.

- Dla pierwszej pochodnej:

$$\overbrace{a_0 w_0 + a_1 w_1 + a_2 w_2}^{\approx w_1^I} = w_1(a_0 + a_1 + a_2) + w_1^I(-h \cdot a_0 + h \cdot a_2) + w_1^{II}(\frac{1}{2} h^2 a_0 + \frac{1}{2} h^2 a_2)$$

$$w_1^I \approx w_1(a_0 + a_1 + a_2) + w_1^I(-h \cdot a_0 + h \cdot a_2) + w_1^{II}(\frac{1}{2} h^2 a_0 + \frac{1}{2} h^2 a_2) \quad ,$$

- Dla drugiej pochodnej:

$$\overbrace{b_0 w_0 + b_1 w_1 + b_2 w_2}^{\approx w_1^{II}} = w_1(b_0 + b_1 + b_2) + w_1^I(-h \cdot b_0 + h \cdot b_2) + w_1^{II}(\frac{1}{2} h^2 b_0 + \frac{1}{2} h^2 b_2)$$

$$w_1^{II} \approx w_1(b_0 + b_1 + b_2) + w_1^I(-h \cdot b_0 + h \cdot b_2) + w_1^{II}(\frac{1}{2} h^2 b_0 + \frac{1}{2} h^2 b_2).$$

Dla obydwu przypadków powstaje układ równań z tą samą macierzą współczynników, ale z innymi prawymi stronami:

$$\begin{bmatrix} 1 & 1 & 1 \\ -h & 0 & h \\ \frac{1}{2}h^2 & 0 & \frac{1}{2}h^2 \end{bmatrix} \cdot \begin{bmatrix} a_0 & b_0 \\ a_1 & b_1 \\ a_2 & b_2 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \Rightarrow \begin{bmatrix} a_0 & b_0 \\ a_1 & b_1 \\ a_2 & b_2 \end{bmatrix} = \begin{bmatrix} -\frac{1}{2h} & \frac{1}{h^2} \\ 0 & \frac{-2}{h^2} \\ \frac{1}{2h} & \frac{1}{h^2} \end{bmatrix}$$

Stąd:

$$w_1^I \approx \frac{1}{2h}(w_2 - w_0)$$

$$w_1^{II} \approx \frac{1}{h^2}(w_0 - 2 \cdot w_1 + w_2)$$

.

Obliczenie dokładności takich wzorów polega na przywróceniu pierwszego z odrzuconych niezerowych wyrazów w każdym z rozwinięć, przemnożeniu przez odpowiedni współczynnik a następnie zsumowaniu.

- Dla pierwszej pochodnej (wyrazy trzeciego rzędu):
  $\varepsilon(h) = a_0 \cdot (-\frac{1}{6}h^3 w_1^{III}) + a_2 \cdot \frac{1}{6}h^3 w_1^{III} = \frac{1}{6}h^3 w_1^{III}(a_2 - a_0) = \frac{1}{6}h^3 w_1^{III}(\frac{1}{2h} + \frac{1}{2h}) = \frac{1}{6}h^2 w_1^{III}$,
- Dla drugiej pochodnej (wyrazy czwartego rzędu):
  $\varepsilon(h) = b_0 \cdot \frac{1}{24}h^4 w_1^{IV} + b_2 \cdot \frac{1}{24}h^4 w_1^{IV} = \frac{1}{24}h^4 w_1^{IV}(b_2 - b_0) = \frac{1}{24}h^4 w_1^{IV}(\frac{1}{h^2} + \frac{1}{h^2}) = \frac{1}{12}h^2 w_1^{IV}$.

Sprawdzenie powyższych wzorów może odbyć się dla wielomianów, dla których wzory dają jeszcze wynik ścisły. W tym przypadku będą to wielomiany rzędu drugiego.

Przyjmijmy funkcję $f(x) = x^2$ oraz następujące węzły:

i	1	2	3
x_i	0	1	2
$f_i = f(x_i)$	0	1	4

Węzły są równooddalone, ich odległość wynosi $h = 1$.

- Ścisłe wartości analityczne pochodnych:
  $f(x) = x^2 \quad \rightarrow \quad f^I(x) = 2x \quad \rightarrow \quad f^{II}(x) = 2$, stąd: $f_1^I = 2, \quad f_1^{II} = 2$.
- Wartości numeryczne pochodnych (ze wzorów różnicowych):
  $w_1^I \approx \frac{4-0}{2 \cdot 1} = 2, \quad w_1^{II} \approx \frac{0 - 2 \cdot 1 + 4}{1^2} = 2$.

Wniosek: $w_1^I = f_1^I = 2, \quad w_1^{II} = f_1^{II} = 2$.

Wyprowadzone wyżej wzory należą do tzw. centralnych wzorów różnicowych. Oprócz nich istnieją też tzw. poboczne wzory różnicowe, o wiele mniej dokładne, np. dla pierwszej pochodnej:

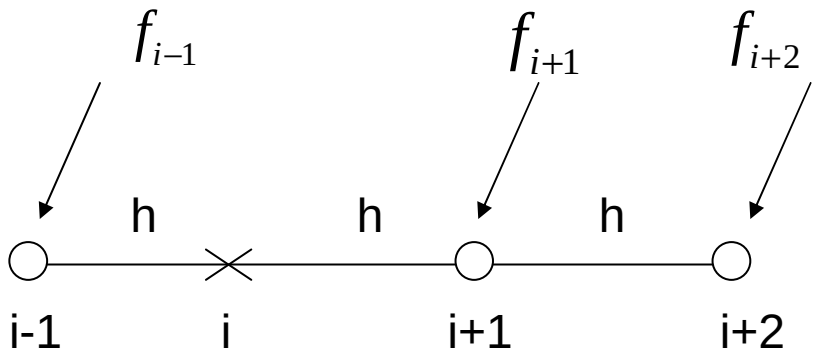
- tzw. iloraz „wprzód”: $w_1^I \approx \frac{w_1 - w_0}{h}$,
- twz. iloraz „wstecz”: $w_1^I \approx \frac{w_2 - w_1}{h}$.

Dają one wyniki ścisłe w ramach pierwszego rzędu wielomianowej aproksymacji. Dla wyżej testowanej funkcji nie dałyby wyników ścisłych, tylko przybliżone.

Przykład 11
Znaleźć przedstawienie operatora drugiej pochodnej w postaci:

$$f_i^{II} = \alpha_i f_{i-1} + \beta_i f_{i+1} + \gamma_i f_{i+2}.$$

Zakładając, iż odstępy między węzłami są stałe i wynoszą *h*, dana konfiguracja węzłów wygląda następująco:



W punkcie (*i*) nie jest dana żadna informacja (a więc nie jest on węzłem), a mimo wszystko poszukuje się w nim wartości numerycznej na drugą pochodną funkcji.

Rozwijamy każdą wartość funkcyjną względem punktu (*i*) w *szereg Taylora* zachowując tyle wyrazów rozwinięcia, ile nieznanych współczynników należy wyznaczyć (3). W takim przypadku uzyskamy interpolację, czyli przeprowadzimy lokalną krzywą paraboliczną przez wszystkie wartości węzłowe.

$$f_{i-1} = f_i - h f_i^I + \frac{1}{2} h^2 f_i^{II}$$
$$f_{i+1} = f_i + h f_i^I + \frac{1}{2} h^2 f_i^{II}$$
$$f_{i+2} = f_i + 2h f_i^I + \frac{1}{2} (2h)^2 f_i^{II}$$

Dalej mnożymy każde z rozwinięć przez niewiadomy współczynnik stojący przy rozwiniętej wartości funkcyjnej we wzorze różnicowym.

$$f_{i-1} = f_i - h f_i^I + \frac{1}{2} h^2 f_i^{II} \quad /\times \alpha_i$$
$$f_{i+1} = f_i + h f_i^I + \frac{1}{2} h^2 f_i^{II} \quad /\times \beta_i$$
$$f_{i+2} = f_i + 2h f_i^I + \frac{1}{2} (2h)^2 f_i^{II} \quad /\times \gamma_i$$

Teraz dodajemy stronami powyższe rozwinięcia, pamiętając o mnożeniu ich przez współczynniki $\alpha_i, \beta_i, \gamma_i$.

$$\underbrace{\alpha_i f_{i-1} + \beta_i f_{i+1} + \gamma_i f_{i+2}}_{\approx f_i^{II}} = f_i(\alpha_i + \beta_i + \gamma_i) + f_i^{I}(-h\cdot\alpha_i + h\cdot\beta_i + 2h\cdot\gamma_i) + f_i^{II}(\frac{1}{2}h^2\alpha_i + \frac{1}{2}h^2\beta_i + 2h^2\gamma_i)$$

Ponieważ wyrażenie lewej strony to wyjściowy wzór różnicowy na drugą pochodną, można zastąpić je wartością drugiej pochodnej.

$$f_i^{II} \approx f_i(\alpha_i + \beta_i + \gamma_i) + f_i^{I}(-h\cdot\alpha_i + h\cdot\beta_i + 2h\cdot\gamma_i) + f_i^{II}(\frac{1}{2}h^2\alpha_i + \frac{1}{2}h^2\beta_i + 2h^2\gamma_i).$$

Aby zachodziła równość między lewą i prawą stroną, współczynniki przy funkcji i jej kolejnych pochodnych muszą być sobie równe.

$$\begin{cases} \alpha_i + \beta_i + \gamma_i = 0 \\ -h\cdot\alpha_i + h\cdot\beta_i + 2h\cdot\gamma_i = 0 \\ \frac{1}{2}h^2\alpha_i + \frac{1}{2}h^2\beta_i + 2h^2\gamma_i = 1 \end{cases}$$

W ten sposób powstaje układ równań na współczynniki $\alpha_i, \beta_i, \gamma_i$. Po jego rozwiązaniu otrzymujemy

$$\begin{cases} \alpha_i = \frac{1}{3h^2} \\ \beta_i = -\frac{1}{h^2} \\ \gamma_i = \frac{2}{3h^2} \end{cases}.$$

Końcowy wzór różnicowy

$$f_i^{II} = \frac{1}{3h^2}(f_{i-1} - 3f_{i+1} + 2f_{i+2})$$

Dokładność wzoru można oszacować zbierając pierwsze odrzucone wyrazy rozwinięć.

$$\varepsilon(h) = -\alpha_i\frac{1}{6}h^3 f_i^{III} + \beta_i\frac{1}{6}h^3 f_i^{III} + \gamma_i\frac{1}{6}(2h)^3 f_i^{III} = \frac{1}{18}h\cdot f_i^{III}(-1-3+8\cdot 2) = \frac{2}{3}h\cdot f_i^{III}.$$

Wzór jest ścisły dla wielomianów rzędu, co najwyżej drugiego. Sprawdzenie będzie polegać na policzeniu pochodnej numerycznej dla funkcji testowej oraz porównanie ze ścisłą wartością. Przyjęto rozstaw węzłów: $x_{i-1} = 0,\ x_{i+1} = 2,\ x_{i+2} = 3$. Rozstaw $h = 1$.

- Funkcja testowa: $f(x) = x^2$
  - Wartości węzłowe: $f_{i-1} = 0^2 = 0,\ f_{i+1} = 2^2 = 4,\ f_{i+2} = 3^2 = 9$.

- Wartość numeryczna drugiej pochodnej: $f_i^{II} \approx \frac{1}{3\cdot 1^2}(0-3\cdot 4+2\cdot 9)=2$.
- Wartość ścisła drugiej pochodnej: $f(x)=x^2 \quad \Rightarrow \quad f^{II}(x)=2 \quad \Rightarrow \quad f_i^{II}=2$.

Numeryczna wartość jest wartością ścisłą. Nie jest to przypadek, gdyż faktycznie dla funkcji parabolicznej $f^{III}(x)=0,$ a więc błąd wyniku $\varepsilon \equiv 0$.

- Funkcja testowa: $f(x)=x^3$
  - Wartości węzłowe: $f_{i-1}=0^2=0,\ f_{i+1}=2^3=8,\ f_{i+2}=3^3=27$.
  - Wartość numeryczna drugiej pochodnej: $f_i^{II} \approx \frac{1}{3\cdot 1^2}(0-3\cdot 8+2\cdot 27)=10$.
  - Wartość ścisła drugiej pochodnej: $f(x)=x^3 \quad \Rightarrow \quad f^{II}(x)=6x \quad \Rightarrow \quad f_i^{II}=6$.

Numeryczna wartość nie jest wartością ścisłą. Błąd wyniku $\varepsilon=\frac{2}{3}\cdot 1\cdot (f_i^{III}=6)=4$ jest w tym przypadku różnicą bezwzględną między wartością numeryczną i ścisłą.

Metoda współczynników nieoznaczonych, oparta na rozwinięciu w *szereg Taylora* ma wiele zalet. Jedną z nich jest możliwość łatwego oszacowania błędu wzoru różnicowego. Metoda pozwala również na budowanie operatorów różniczkowych dowolnej postaci, np.

$$\mathcal{L}=a\cdot\frac{d^2}{dx^2}+b\cdot\frac{d}{dx}+c, \quad a,b,c\in\Re$$

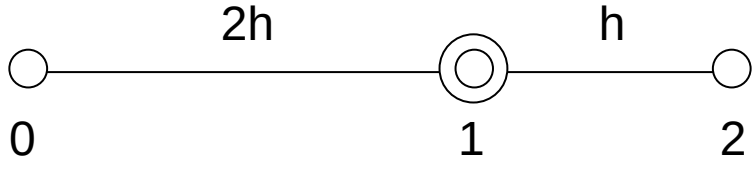
poprzez przybliżanie ich wartości w węźle (i) wzorem interpolacyjnym opartym na trzech węzłach:

$$\mathcal{L}u_i \approx Lu_i=\alpha_{i-1}u_{i-1}+\beta_i u_i+\gamma_{i+1}u_{i+1}.$$

Inną cechą tak budowanych wzorów różnicowych jest to, iż mogą one bazować nie tylko na wartościach samej funkcji w węzłach, ale także ich kolejnych pochodnych (byle nie wyższych niż najwyższy rząd pochodnej występującej w operatorze różniczkowym). Wartości pochodnych funkcji w węzłach (lub nawet wartości całych operatorów różniczkowych) nazywane są uogólnionymi stopniami swobody.

Przykład 12

Znaleźć numeryczną wartość operatora różniczkowego $\mathcal{L}u_1=-2u_1+4u_1^I-3u_1^{II}$ za pomocą następującego wzoru różnicowego $Lu_1=\alpha u_0+\beta u_1+\gamma u_2^I$ dla zadanej konfiguracji węzłów jak na rys. Wzór sprawdzić dla funkcji testowych $x^2,\ x^3$. Określić dokładność takiego wzoru.



Rozwinięcie wartości węzłowych w *szereg Taylora* i przemnożenie rozwinięć przez odpowiedni współczynnik:

$$u_0 = u_1 - 2h\cdot u_1^I + \frac{1}{2}(-2h)^2 u_1^{II} + \ldots \qquad /\times\alpha$$
$$u_1 = u_1 \qquad /\times\beta$$
$$u_2^I = u_1^I + h\cdot u_1^{II} + \ldots \qquad /\times\gamma$$

Ostatnie równanie to rozwinięcie wartości pierwszej pochodnej. Znajduje się je poprzez rozwinięcie samej wartości funkcji: $u_2 = u_1 + h\cdot u_1^I + \frac{1}{2}h^2\cdot u_1^{II} + \ldots$ a następnie różniczkuje się je stronami (tak, aby otrzymać po stronie lewej pierwszą pochodną) opuszczając wyrazy rzędu wyższego niż drugi.

Dodanie rozwinięć stronami:

$$\alpha\cdot u_0 + \beta\cdot u_1 + \gamma\cdot u_2^I = u_1(\alpha+\beta) + u_1^I(-2h\cdot\alpha+\gamma) + u_1^{II}(2h^2\cdot\alpha + h\cdot\gamma)$$

oraz zastąpienie (w przybliżeniu) wzoru różnicowego (lewa strona) wartością operatora różniczkowego:

$$\underbrace{\alpha\cdot u_0 + \beta\cdot u_1 + \gamma\cdot u_2^I}_{\mathcal{L}u_1 = -2u_1 + 4u_1^I - 3u_1^{II}} = u_1(\alpha+\beta) + u_1^I(-2h\cdot\alpha+\gamma) + u_1^{II}(2h^2\cdot\alpha + h\cdot\gamma) \quad \Rightarrow$$
$$\Rightarrow \qquad -2u_1 + 4u_1^I - 3u_1^{II} \cong u_1(\alpha+\beta) + u_1^I(-2h\cdot\alpha+\gamma) + u_1^{II}(2h^2\cdot\alpha + h\cdot\gamma)$$

prowadzi, po przyrównaniu współczynników przy funkcji i odpowiednich pochodnych do końcowego układu równań algebraicznych:

$$\begin{cases} \alpha+\beta = -2 \\ -2h\cdot\alpha+\gamma = 4 \\ 2h^2\cdot\alpha + h\cdot\gamma = -3 \end{cases} \quad \Rightarrow \quad \begin{cases} \alpha = \dfrac{-3-4h}{4h^2} \\ \beta = \dfrac{-8h^2+3+4h}{4h^2} \\ \gamma = \dfrac{-3+4h}{2h} \end{cases}$$

Końcowa postać wzoru różnicowego:

$$\mathcal{L}u_1 = -2u_1 + 4u_1^I - 3u_1^{II} \approx Lu_1 = \frac{-3-4h}{4h^2}u_0 + \frac{-8h^2+3+4h}{4h^2}u_1 + \frac{-3+4h}{2h}u_2 .$$

Dokładność wzoru:

$$\varepsilon(h) = \frac{1}{6}(-2h)^3\cdot u_1^{III}\cdot\alpha + \frac{1}{2}h^2 u_1^{III}\cdot\gamma = \frac{h}{12}u_1^{III}(3+28h) .$$

Sprawdzenie dla jednomianów:
(przyjęto: $x_0 = 1, \quad x_1 = 3, \quad x_2 = 4 \quad \rightarrow \quad h = 1$)

- dla $u(x) = x^2 \ (u^I(x) = 2x, \quad u^{II}(x) = 2, \quad u^{III}(x) = 0)$

  Wartość ścisła: $u_1 = 9, u_1^I = 6, u_1^{II} = 2 \quad \rightarrow \quad \mathcal{L}u_1 = -2 \cdot 9 + 4 \cdot 6 - 3 \cdot 2 = 0$
  Wartość numeryczna:
  $u_0 = 1, u_1 = 9, u_2^I = 8 \quad \rightarrow \quad Lu_1 = \frac{-3-4}{4} \cdot 1 + \frac{3+4-8}{4} \cdot 9 + \frac{4-3}{2} \cdot 8 = 0.$
  Błąd wyniku: $\varepsilon = \frac{1}{12} \cdot 0 \cdot (3+28) = 0.$

- dla $u(x) = x^3 \ (u^I(x) = 3x^2, \quad u^{II}(x) = 6x, \quad u^{III}(x) = 6)$

  Wartość ścisła: $u_1 = 27, u_1^I = 27, u_1^{II} = 18 \quad \rightarrow \quad \mathcal{L}u_1 = -2 \cdot 27 + 4 \cdot 27 - 3 \cdot 18 = 0$
  Wartość numeryczna:
  $u_0 = 1, u_1 = 27, u_2^I = 48 \quad \rightarrow \quad Lu_1 = \frac{-3-4}{4} \cdot 1 + \frac{3+4-8}{4} \cdot 27 + \frac{4-3}{2} \cdot 48 = 15.5.$
  Błąd wyniku: $\varepsilon = \frac{1}{12} \cdot 6 \cdot (3+28) = 15.5$

Operatory różnicowe można też budować metodami aproksymacji funkcji dyskretnej, np. *najlepszej aproksymacji*. Wyniki mogą się różnić od wyników pochodzących z interpolacji (zwłaszcza, jeżeli w tzw. *gwieździe*, – czyli konfiguracji węzłów – jest nadmiar węzłów w stosunku do niezbędnej liczby informacji potrzebnej do zbudowania odpowiedniego operatora). Techniką powszechnie używaną w metodach dyskretnych do rozwiązywania równań różniczkowych brzegowych (zwłaszcza w bezsiatkowej metodzie różnic skończonych *BMRS*) służącą do generacji kompletów wzorów różnicowych jest technika aproksymacji *MWLS* (ang. *Moving Weighted Least Squares*) – *technika najmniejszych ważonych kroczących kwadratów*.

III. NUMERYCZNE CAŁKOWANIE FUNKCJI

Tak jak wynikiem numerycznego różniczkowania była wartość dowolnej pochodnej w konkretnym węźle (lub w dowolnym punkcie), tak wynikiem całkowania numerycznego nie jest funkcja analityczna, a jedynie wartość liczbowa całki. Stąd oczywisty wniosek, iż numeryka pozwala na obliczanie przede wszystkim całek oznaczonych (liczb) w dowolnym przedziale (a,b). Wzory całkowania numerycznego, zwane kwadraturami, pozwalają na obliczenie (w przybliżeniu) wartości całki:

$$I = \int_a^b f(x)dx$$

Na początek zakładamy, iż granice całkowania są skończone, a funkcja podcałkowa nie ma w przedziale (a,b) osobliwości (jest ciągła) – tzw. całka właściwa. Wzory te dzielimy na dwie główne grupy:

- *kwadratury Newtona – Cotesa*, polegające na zastąpieniu funkcji podcałkowej wielomianami coraz to wyższych rzędów w przedziale podzielonym na odcinki równej długości,
- *kwadratury Gaussa*, polegające na zastąpieniu funkcji podcałkowej wielomianami ortogonalnymi w taki sposób, aby wzór był ścisły dla wielomianu możliwie najwyższego rzędu.

Po zastąpieniu funkcji podcałkowej wielomianem, łatwym do scałkowania, otrzymujemy wzór całkowania, bazujący na wartościach funkcji w przedziale (a,b).

Kwadratury Newtona – Cotesa

Funkcja podcałkowa jest aproksymowana przez wielomiany coraz to wyższych rzędów w przedziale (a,b) podzielonym na odcinki o równej długości (podział równomierny).

Założenie: $x_{i+1} - x_i = x_i - x_{i-1} = h = const.$

Przedział (a,b) dzielimy na podprzedziały o równej długości punktami $x_i, \quad i = 0,1,2,...,n$, $a = x_0 + ph, \quad b = x_0 + qh, \quad p \geq 0, \; q \leq n$.

Budujemy *wielomiany Lagrange'a*:

$$I = \int_a^b f(x)dx \approx \int_a^b \sum_{j=0}^{n} L_j^{(n)}(x) f_j dx = I_{(n+1)} = \sum_{j=0}^{n} \int_a^b L_j^{(n)}(x) f_j dx .$$

Wprowadzamy indeks *s* tak, że $x = x_0 + sh$.

$$I_{(n+1)} = \sum_{j=0}^{n} \int_p^q \prod_{\substack{k=0 \\ k\neq j}}^{n} \frac{(x_0+sh)-(x_0+kh)}{(x_0+jh)-(x_0+kh)} f_j h ds = h \sum_{j=0}^{n} \int_p^q \prod_{\substack{k=0 \\ k\neq j}}^{n} \frac{s-k}{j-k} f_j ds = h \sum_{j=0}^{n} f_j \prod_{\substack{k=0 \\ k\neq j}}^{n} \frac{1}{j-k} \int_p^q \prod_{k=0}^{n} \frac{s-k}{s-j} ds$$

Wprowadzamy współczynniki liczbowe α_j

$$\alpha_j = \frac{h}{\prod_{\substack{k=0 \\ k\neq j}}^{n} \frac{1}{j-k} \int_p^q \prod_{k=0}^{n} \frac{s-k}{s-j} ds} = h \frac{(-1)^{n-j}}{n!} \binom{n}{j} \int_p^q \frac{\prod_{k=0}^{n} s-k}{s-j} ds, \quad j = 0,1,...,n .$$

Ostateczna postać kwadratury

$$I = \int_a^b f(x)dx \approx \sum_{j=0}^{n} \alpha_j f_j + E, \quad E = I - I_{(n+1)} ,$$

gdzie błąd *E* wyniku wyraża się wzorem:

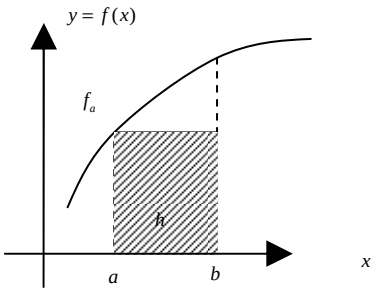
$$E=\begin{cases}\dfrac{2h^{n+2}}{(n+1)!}f^{(n+1)}(\xi)\displaystyle\int_0^{m+\frac{1}{2}}\prod_{k=\frac{-2r+1}{2}}^{k=\frac{2r-1}{2}}(s-k)ds, & \text{dla } n \text{ nieparzystych } (\xi\in\langle a,b\rangle)\\ \dfrac{2h^{n+3}}{(n+2)!}f^{(n+2)}(\eta)\displaystyle\int_0^{m}\prod_{k=0}^{k=r}(s^2-k^2)ds, & \text{dla } n \text{ parzystych } (\eta\in\langle a,b\rangle)\end{cases}$$

Tabela współczynników $\frac{\alpha_i}{h}$ wzorów Newtona – Cotesa

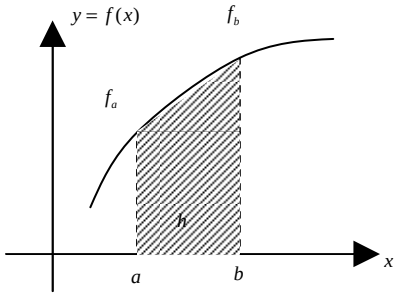
n	$j=0$	$j=1$	$j=2$	$j=3$	*błąd*	*nazwa wzoru*
1	$\frac{1}{2}$	$\frac{1}{2}$	—	—	$-\frac{1}{12}h^3 f^{II}(\xi)$	*wzór trapezów*
2	$\frac{1}{3}$	$\frac{4}{3}$	$\frac{1}{3}$	—	$-\frac{1}{90}h^5 f^{IV}(\xi)$	*wzór Simpsona*
3	$\frac{3}{8}$	$\frac{9}{8}$	$\frac{9}{8}$	$\frac{3}{8}$	$-\frac{3}{80}h^5 f^{IV}(\xi)$	—

Szczególnie korzystny w zastosowaniach jest *wzór Simpsona* ze względu na podwyższoną dokładność.

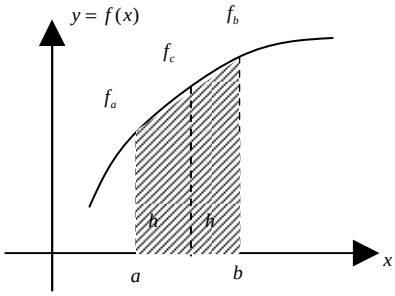
Trzy pierwsze *kwadratury Newtona – Cotesa* to najpowszechniej używane wzory całkowania numerycznego.



Wzór prostokątów



Wzór trapezów



Wzór Simpsona

- *Wzór prostokątów*

$$I=\int_a^b f(x)dx\approx\int_a^b f(a)dx=f_a\,x\Big|_a^b=f_a(b-a)=f_a h\,.$$

- *Wzór trapezów*

$$I=\int_a^b f(x)dx\approx\int_a^b f(a)\frac{h-x}{x}+f(b)\frac{x}{h}dx=\frac{h}{2}(f_a+f_b)\,.$$

- *Wzór Simpsona*

$$I=\int_a^b f(x)dx \approx \int_a^b f(a)L_0^{(2)}(x)+f(\frac{a+b}{2})L_1^{(2)}(x)+f(b)L_2^{(2)}(x)dx=\frac{h}{3}(f_a+4\cdot f_c+f_b)$$

$$c=\frac{a+b}{2},\quad h=\frac{b-a}{2}$$

W praktyce nie używa się już wzorów wyższego rzędu, natomiast stosuje się powyższe trzy wzory niskich rzędów (zwłaszcza *wzór Simpsona*) w podprzedziałach wynikających z podziału wyjściowego przedziału (a,b). Powstają w ten sposób tzw. *wzory złożone całkowania*. Ilość podziałów nie jest z góry założona, należy ją dobrać iteracyjnie ze względu na żądaną dokładność wyników ε.

Przykład 12

Podaną całkę $I=\int_0^1 \sqrt{1+x}\,dx$ obliczyć numerycznie stosując *wzory Newtona – Cotesa* proste i złożone (dwa podziały). Za każdym razem porównać otrzymany wynik numeryczny z rozwiązaniem analitycznym.

Wynik analityczny

$$I=\int_0^1 \sqrt{1+x}\,dx=\frac{2}{3}(1+x)^{\frac{3}{2}}\Big|_0^1=\frac{2}{3}\cdot 2^{\frac{3}{2}}-\frac{2}{3}=1.218951.$$

Proste wzory całkowania (1 przedział)

$$I_{p_1}=f(a)\cdot h=\sqrt{1+0}\cdot 1=1,\quad \varepsilon=\left|\frac{I_{p_1}-I}{I}\right|\cdot 100\%=\left|\frac{1-1.218951}{1.218951}\right|\cdot 100\%=18\%$$

$$I_{t_1}=\frac{f(a)+f(b)}{2}\cdot h=\frac{1}{2}\cdot 1\cdot(\sqrt{1+0}+\sqrt{1+1})=\frac{1+\sqrt{2}}{2}=1.207107,$$

$$\varepsilon=\left|\frac{I_{t_1}-I}{I}\right|\cdot 100\%=\left|\frac{1.207107-1.218951}{1.218951}\right|\cdot 100\%=1\%$$

$$I_{S_1}=\frac{b-a}{6}(f(a)+4f(c)+f(b))=\frac{1}{6}\cdot(\sqrt{1+0}+4\sqrt{1+\frac{1}{2}}+\sqrt{1+1})=\frac{1+4\sqrt{\frac{3}{2}}+\sqrt{2}}{6}=1.218866,$$

$$\varepsilon=\left|\frac{I_{S_1}-I}{I}\right|\cdot 100\%=\left|\frac{1.218866-1.218951}{1.218951}\right|\cdot 100\%=0.007\%$$

Złożone wzory całkowania (2 równe przedziały, $h=\frac{1}{2}$)

$$I_{p_2}=f(0)\cdot\frac{1}{2}+f(1)\cdot\frac{1}{2}=\sqrt{1+0}\cdot\frac{1}{2}+\sqrt{1+\frac{1}{2}}\cdot\frac{1}{2}=\frac{1}{2}+\frac{1}{2}\sqrt{\frac{3}{2}}=1.112372,$$

$$\varepsilon=\left|\frac{I_{p_2}-I}{I}\right|\cdot 100\%=\left|\frac{1.112372-1.218951}{1.218951}\right|\cdot 100\%=8.7\%$$

$$I_{t_2} = \left[f(0) + f(\frac{1}{2}) \right] \cdot \frac{1}{2} \cdot \frac{1}{2} + \left[f(\frac{1}{2}) + f(1) \right] \cdot \frac{1}{2} \cdot \frac{1}{2} = \left(\sqrt{1+0} + \sqrt{1+\frac{1}{2}} \right) \cdot \frac{1}{4} + \left(\sqrt{1+\frac{1}{2}} + \sqrt{1+1} \right) \cdot \frac{1}{4} =$$

$$= \frac{1}{4}(1 + 2\sqrt{\frac{3}{2}} + \sqrt{2}) = 1.215926, \quad \varepsilon = \left| \frac{I_{t_2} - I}{I} \right| \cdot 100\% = \left| \frac{1.215926 - 1.218951}{1.218951} \right| \cdot 100\% = 0.2\%$$

$$I_{S_2} = \left[f(0) + 4f(\frac{1}{4}) + f(\frac{1}{2}) \right] \cdot \frac{1}{4 \cdot 3} + \left[f(\frac{1}{2}) + 4f(\frac{3}{4}) + f(1) \right] \cdot \frac{1}{4 \cdot 3} = \left(1 + 4\sqrt{\frac{5}{4}} + 2\sqrt{\frac{3}{2}} + 4\sqrt{\frac{7}{4}} + \sqrt{2} \right) \cdot \frac{1}{12} =$$

$$= 1.218945, \quad \varepsilon = \left| \frac{I_{t_2} - I}{I} \right| \cdot 100\% = \left| \frac{1.218945 - 1.218951}{1.218951} \right| \cdot 100\% = 0.0005\%$$

Kwadratury Gaussa

We *wzorach Gaussa* zastępujemy całkę analityczną w przedziale $\langle -1,1 \rangle$ kombinacją liniową wartości funkcji podcałkowej $f(x_i)$ w tzw. *punktach Gaussa* x_i (węzły całkowania) oraz wag liczbowych ω_i.

$$\int_{-1}^{1} f(x)dx \approx \sum_{i=1}^{N} \omega_i f(x_i)$$

N oznacza ilość *punktów Gaussa* (jak i również wag).

Wagi i węzły całkowania ustala się według zasady, by wzór całkowania przybliżony był wzorem ścisłym dla wielomianu możliwie wysokiego stopnia.

$$\sum_{i=1}^{N} \omega_i f(x_i) \approx \sum_{i=1}^{N} \omega_i x_i^k = \int_{-1}^{1} x^k dx = \frac{1}{k+1} \left[1 - (-1)^{k+1} \right], \quad k = 0,1,2,...,2N-1.$$

Np. dla $N = 2$ (*wzór dwupunktowy Gaussa*)

$$\begin{aligned} k = 0 \quad &\rightarrow \quad \omega_1 \cdot 1 + \omega_2 \cdot 1 = 2 \\ k = 1 \quad &\rightarrow \quad \omega_1 \cdot x_1 + \omega_2 \cdot x_2 = 0 \\ k = 2 \quad &\rightarrow \quad \omega_1 \cdot x_1^2 + \omega_2 \cdot x_2^2 = \frac{2}{3} \\ k = 3 \quad &\rightarrow \quad \omega_1 \cdot x_1^3 + \omega_2 \cdot x_2^3 = 0 \end{aligned} \quad \Rightarrow \quad \begin{cases} \omega_1 = \omega_2 = 1 \\ x_1 = -\frac{1}{\sqrt{3}}, \quad x_2 = \frac{1}{\sqrt{3}} \end{cases}.$$

W praktyce wag i *punktów Gaussa* nie znajduje się w powyższego warunku. Pochodzą one mianowicie od rodziny pewnych wielomianów ortogonalnych (z wagą) w przedziale $\langle -1,1 \rangle$. Wtedy *punkty Gaussa* są ich miejscami zerowymi. Powyższe liczby pochodzą od tzw. *wielomianów Legendre'a* – wtedy *wzory Gaussa* nazywane są *wzorami Gaussa – Legendre'a*. Wartości wag i miejsc zerowych tych wielomianów są tablicowane, tak jak innych kwadratur wykorzystujących wielomiany ortogonalne.

Tablica rodziny *wzorów Gaussa – Legendre'a*

Stopień *wielomianu Legendre'a*	Miejsca zerowe *wielomianów Legendre'a* x_i	Wagi ω_i
1	0	2
2	$\pm\frac{1}{\sqrt{3}}$	1, 1
3	$0, \pm\sqrt{\frac{3}{5}}$	$\frac{8}{9}, \frac{5}{9}, \frac{5}{9}$
4	±0.3399810436 ±0.8611363116	0.6521451549 0.3478548451

W ogólnym przypadku liczymy całkę z dowolnego przedziału (a,b). Konieczna jest więc transformacja liniowa między danym przedziałem a przedziałem $\langle -1,1\rangle$, tak aby można było zastosować powyższe dane z tabeli, obowiązujące tylko w tym przedziale.

Niech $I=\int_a^b f(z)dz, \quad z\in(a,b)$.

Wzory na transformację $(a,b) \;\Rightarrow\; \langle -1,1\rangle$

$$\begin{matrix} x=\frac{2z-(a+b)}{b-a} \\ dx=\frac{2}{b-a}dz \end{matrix} \quad \rightarrow \quad \begin{matrix} z=\frac{b-a}{2}x+\frac{a+b}{2} \\ dz=\frac{b-a}{2}dx \end{matrix}$$

$$I=\int_a^b f(z)dz=\int_{-1}^{1} f(z(x))\frac{dz}{dx}dx=\frac{b-a}{2}\int_{-1}^{1} f(\frac{b-a}{2}x+\frac{a+b}{2})dx=\frac{b-a}{2}\int_{-1}^{1} f(x)dx\approx\sum_{i=1}^{N}\omega_i f(x_i)$$

Przykład 13

Obliczyć całkę z poprzedniego przykładu $I=\int_0^1\sqrt{1+z}\,dz$ stosując *wzory Gaussa*: dwupunktowy i trzypunktowy.

$$a=0, \quad b=1 \;\rightarrow\; \begin{cases} z=\frac{1-0}{2}x+\frac{1+0}{2}=\frac{1}{2}x+\frac{1}{2} \\ dx=\frac{1}{2}dx \end{cases}$$

$$I=\int_0^1\sqrt{1+z}\,dz=\frac{1}{2}\int_{-1}^{1}\sqrt{\frac{1}{2}x+\frac{3}{2}}dx.$$

Wzór dwupunktowy:

$$I_{G_2} = \frac{1}{2}\left[1\cdot\sqrt{\frac{1}{2}\cdot\frac{1}{\sqrt{3}}+\frac{3}{2}}+1\cdot\sqrt{\frac{1}{2}\cdot\left(-\frac{1}{\sqrt{3}}\right)+\frac{3}{2}}\right] = 1.219008,$$

$$\varepsilon = \left|\frac{I_{G_2}-I}{I}\right|\cdot 100\% = \left|\frac{1.219008-1.218951}{1.218951}\right|\cdot 100\% = 0.005\%$$

.

Wzór trzypunktowy:

$$I_{G_3} = \frac{1}{2}\left[\frac{8}{9}\cdot\sqrt{0+\frac{3}{2}}+\frac{5}{9}\cdot\sqrt{\frac{1}{2}\cdot\sqrt{\frac{3}{5}}+\frac{3}{2}}+\frac{5}{9}\cdot\sqrt{\frac{1}{2}\cdot\left(-\sqrt{\frac{3}{5}}\right)+\frac{3}{2}}\right] = 1.218952,$$

$$\varepsilon = \left|\frac{I_{G_3}-I}{I}\right|\cdot 100\% = \left|\frac{1.218952-1.218951}{1.218951}\right|\cdot 100\% = 0.00007\%$$

.

Wszystkie omawiane powyżej wzory całkowania numerycznego dotyczyły przypadków nieosobliwych, tzn. tzw. całej właściwych. Istnieją też całki niewłaściwe, gdy jedna z granic całkowania to nieskończoność (niewłaściwość I rodzaju) lub istnieje osobliwość funkcji podcałkowej w jednej z granic (niewłaściwość II rodzaju). W tych przypadkach na ogół nie da się stosować bezpośrednio wzorów całkowania numerycznego, należy dodatkowo przekształcić całkę analitycznie. W przypadku nieskończoności w jednej z granic wzory Newtona i Gaussa są bezużyteczne, bo nie da się wprowadzić węzłów całkowania do przedziału nieskończonego. W drugim przypadku osobliwości funkcji podcałkowej w jednej z granic nie można stosować wzorów Newtona – gdyż wymagają one znajomości wartości funkcji podcałkowej w jednej z granic – a jest ona równa nieskończoności. Wzory Gaussa można stosować, gdyż węzły całkowania pochodzą wtedy z wnętrza przedziału i nie natrafiają na punkt osobliwy.

Całki niewłaściwe I rodzaju

Można je przedstawić w postaci ogólnej $\int_a^{\infty} f(x)dx$.

Analityczne rozwiązanie wymaga liczenia granicy $\int_a^{\infty} f(x)dx = F(x)\Big|_a^{\infty} = \lim_{x\to\infty} F(x) - F(a)$.

Numeryczne rozwiązanie wymaga podstawienia typu $t = \frac{1}{x}$. Wtedy korzystając z twierdzenia o zmianie granic otrzymuje się nowe, skończone granice całkowania $t_1 = t(\infty) = \frac{1}{\infty} = 0, \quad t_2 = t(a) = \frac{1}{a}$. Postać całki nadaje się już do całkowania numerycznego.

$$I = \int_a^{\infty} f(x)dx = \int_{\frac{1}{a}}^{0} f(x(t))\frac{dx}{dt}dt = \ldots$$

Wyjątkowo „złośliwa” jest następująca całka osobliwa $I=\int_0^{\infty} f(x)dx$. Proponowane podstawienie nie odniesie żądanego skutku, dlatego iż na powrót dostaniemy granicę całkowania równą nieskończoności $t_1=t(\infty)=\frac{1}{\infty}=0,\quad t_2=t(a=0)=\frac{1}{0}=\infty(!)$. Dlatego też należy np. rozłożyć całkę na dwie całki składowe, tak, aby całka niewłaściwa miała drugą granicę różną od zera.

$$I=\int_0^{\infty} f(x)dx=\int_0^{1} f(x)dx+\int_1^{\infty} f(x)dx\,.$$

Pierwszą całkę obliczamy numerycznie bezpośrednio, drugą składową po opisanym wyżej podstawieniu.

Całki niewłaściwe II rodzaju

Ogólna postać całki: $I=\int_a^b \frac{f(x)}{(x-a)^k}dx,\quad k\in\Re,\quad k\neq 0$.

Aby pozbyć się osobliwości, należy usunąć ją z mianownika funkcji podcałkowej. Można to zrobić również przez podstawienie, ale łatwiejsze będzie w tym przypadku zastosowanie twierdzenia o całkowaniu przez części.

Całkowanie przez części: $I=\int_a^b f(x)g'(x)dx=\begin{vmatrix} f(x) & f'(x) \\ g'(x) & g(x)\end{vmatrix}=\left[f(x)g(x)\right]_a^b-\int_a^b f'(x)g(x)$.

W omawianym przypadku dla $k=\frac{1}{2}$

$$I=\int_a^b \frac{f(x)}{\sqrt{x-a}}dx=\begin{vmatrix} f(x) & f'(x) \\ g(x)=\frac{1}{\sqrt{x-a}} & 2\sqrt{x-a}\end{vmatrix}=\left[2f(x)\sqrt{x-a}\right]_a^b-2\int_a^b f'(x)\sqrt{x-a}dx\,.$$

Ostatnia całkę można policzyć numerycznie bez żadnych trudności. Dla innych wartości *k* należy powtórzyć całkowanie przez części tak, aby otrzymać w końcu całkę właściwą. Przypadek szczególny $k=1$ doprowadzi do funkcji logarytmicznej, która będzie miała znowu osobliwość dla $x=a$. Taką całkę należy obliczać *kwadraturami Gaussa*.

Przykład 14

Obliczyć numerycznie następujące całki niewłaściwe $I=\int_a^{\infty}\frac{dz}{z^2}$ oraz $I=\int_0^1\frac{x}{\sqrt{1-x}}$

Wyniki analityczne

$$I=\int_1^{\infty}\frac{dz}{z^2}=\left.\frac{z^{-1}}{-1}\right|_1^{\infty}=\left.-\frac{1}{z}\right|_1^{\infty}=(-0+1)=1\,.$$

$$I=\int_0^1\frac{x}{\sqrt{1-x}}=-\int_0^1\frac{\sqrt{(1-x)^2}-1}{\sqrt{1+x}}dx=-\int_0^1\sqrt{1-x}dx+\int_0^1\frac{1}{\sqrt{1-x}}dx=\left[-\frac{2}{3}(1-x)^{\frac{3}{2}}+2(1-x)^{\frac{1}{2}}\right]_0^1=1.333333$$

Przekształcenia analityczne (dla obliczeń numerycznych)

$$I=\int_1^{\infty}\frac{dz}{z^2}=\left|\begin{array}{ll} t=\frac{1}{z^2}\to z=\frac{1}{\sqrt{t}} & t(\infty)=0 \\ dz=-\frac{1}{2}t^{-\frac{3}{2}}dt & t(1)=1 \end{array}\right|=\int_1^0 t\cdot(-\frac{1}{2})\cdot t^{-\frac{3}{2}}dt=\frac{1}{2}\int_0^1\frac{dt}{\sqrt{t}}.$$

$$I=\int_0^1\frac{x}{\sqrt{1-x}}=\left|\begin{array}{ll} f(x)=x & f'(x)=1 \\ g'(x)=\frac{1}{\sqrt{1-x}} & 2\sqrt{1-x} \end{array}\right|=\underbrace{(2x\sqrt{1-x})_0^1}_{0}-2\int_0^1\sqrt{1-x}dx=-2\int_0^1\sqrt{1-x}dx.$$

Obliczenia numeryczne (*wzór* dwupunktowy *Gaussa*)

$$I=\frac{1}{2}\int_0^1\frac{dt}{\sqrt{t}}=\left|\begin{array}{l} t(x)=\frac{1}{2}x+\frac{1}{2} \\ dt=\frac{1}{2}dx \end{array}\right|=\frac{1}{4}\int_{-1}^1\frac{dx}{\sqrt{\frac{1}{2}x+\frac{1}{2}}}\approx\frac{1}{4}(1\cdot\frac{1}{\sqrt{\frac{1}{2}\cdot\frac{1}{\sqrt{3}}+\frac{1}{2}}}+1\cdot\frac{1}{\sqrt{\frac{1}{2}\cdot\left(-\frac{1}{\sqrt{3}}\right)+\frac{1}{2}}})=0.825340$$

$$\varepsilon=\left|\frac{I_{G_2}-I}{I}\right|\cdot 100\%=\left|\frac{0.825340-1.0}{1.0}\right|\cdot 100\%=17.4\%$$

$$I=-2\int_0^1\sqrt{1-t}dt=2\int_1^0\sqrt{1-t}dt=\left|\begin{array}{l} t(x)=-\frac{1}{2}x+\frac{1}{2} \\ dt=-\frac{1}{2}dx \end{array}\right|=\int_{-1}^1\sqrt{\frac{1}{2}+\frac{1}{2}x}dx\approx 1\cdot\sqrt{\frac{1}{2}+\frac{1}{2}\cdot\frac{1}{\sqrt{3}}}+1\cdot\sqrt{\frac{1}{2}+\frac{1}{2}\cdot\left(-\frac{1}{\sqrt{3}}\right)}=1.347775$$

$$\varepsilon=\left|\frac{I_{G_2}-I}{I}\right|\cdot 100\%=\left|\frac{1.347775-1.333333}{1.333333}\right|\cdot 100\%=1.1\%$$

IV. NUMERYCZNE ROZWIĄZYWANIE PROBLEMÓW POCZĄTKOWYCH

Ogólne sformułowanie problemu początkowego

$$\frac{d^{(n)}y}{dx^{(n)}}=f(x,y,y',...,y^{(n-1)}),\quad x\in(a,b)$$

$$y(x_0)=y_0,\quad y'(x_0)=y'_0,\quad ...,\quad y^{(n-1)}(x_0)=y_0^{(n-1)};\quad x_0\in(a,b)$$

Szczególnym przypadkiem problemu początkowego jest równanie różniczkowe rzędu pierwszego z warunkiem na niewiadomą funkcję. Równania wyższych rzędów sprowadza się do równania rzędu pierwszego i rozwiązuje niezależnie.

$$\frac{dy}{dx}=f(x,y),\quad x\in(a,b),\quad y(x_0)=y_0;\quad x_0\in(a,b)$$

Metody numeryczne pozwalają na wyznaczenie zbioru wartości dyskretnych funkcji niewiadomej $y=y(x)$ począwszy od punktu początkowego x_0. Zbiór par(x_i,y_i) wyznacza się z następujących zależności (dla węzłów równoodległych $x_{i+1}-x_i=x_i-x_{i-1}=h=const.$)

$$\begin{cases} x_{i+1} = x_i + h = x_0 + i \cdot h \\ y_{i+1} = y_i + \int_{x_i}^{x_{i+1}} f(t,y)dt = y_i + \Delta y_i, \quad y_0 = y(x_0) \end{cases}$$

Całkę oznaczoną przez Δy_i oblicza się numerycznie na różne sposoby. W zależności od sposobu jej obliczania metody numeryczne do rozwiązywania zadań początkowych można podzielić na jednokrokowe $\Delta y_i = \Delta y_i(f_i), \quad f_i \equiv f(x_i, y_i)$ (wartość delty zależy tylko od jednego punktu wstecz) i wielokrokowe $\Delta y_i = \Delta y_i(f_i, f_{i-1}, f_{i-2}, ...)$ (wartość delty zależy od kilku punktów wstecz). Inna klasyfikacja dotyczy tzw. jawności metod. Przedstawione wyżej wzory dotyczyły metod jawnych (otwartych, ekstrapolacyjnych) – wartość y_{i+1} liczona jest na podstawie znanych wartości funkcji danych lub obliczonych wcześniej w poprzednich punktach - $\Delta y_i = \Delta y_i(f_i, f_{i-1}, f_{i-2}, ...)$. Natomiast inną grupę metod stanowią bardzo dokładne metody niejawne (zamknięte, interpolacyjne), gdzie wartość y_{i+1} zależna jest od siebie samej poprzez deltę $\Delta y_i = \Delta y_i(f_{i+1}, f_i, f_{i-1}, ...)$. Oblicza się ją stosując metody iteracyjne, startujące ze wstępnego określenia wartości $y_{i+1}^{(0)}$ znanego z metody jedno- lub wielokrokowej otwartej.

Metody jednokrokowe

- *Metoda Eulera* (metoda ta zakłada stałość funkcji *y(x)* na odcinku (x_i, x_{i+1})).

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

- *Metoda ulepszona Eulera*

  $$\begin{cases} y_{i+1} = y_i + h \cdot f(x_i, y_i) \\ \tilde{f}_i = \dfrac{f_i + f_{i+1}}{2} \\ y_{i+1} = y_i + h \cdot \tilde{f}_i \end{cases}$$

- *Metoda Rungego – Kutty II rzędu*

  $$K_1 = h \cdot f(x_i, y_i)$$
  $$K_2 = h \cdot f(x_i + h, y_i + K_1)$$
  $$y_{i+1} = y_i + \frac{1}{2}(K_1 + K_2)$$

- *Metoda Rungego – Kutty IV rzędu*

  $$K_1 = h \cdot f(x_i, y_i)$$
  $$K_2 = h \cdot f(x_i + \frac{1}{2}h, y_i + \frac{1}{2}K_1)$$

$$K_3 = h \cdot f(x_i + \frac{1}{2}h, y_i + \frac{1}{2}K_2)$$
$$K_4 = h \cdot f(x_i + h, y_i + K_3)$$
$$y_{i+1} = y_i + \frac{1}{6}(K_1 + 2K_2 + 2K_3 + K_4)$$

Metody wielokrokowe

- *Metoda Adamsa – Bashfortha* (metoda otwarta)

$$y_{i+1} = y_i + h \cdot \sum_{j=n}^{j=0} f_{i-j} \cdot L_{i-j}^{(n)} = y_i + h \cdot (f_{i-n} \cdot L_{i-n}^{(n)} + ... + f_{i-1} \cdot L_{i-1}^{(n)} + f_i \cdot L_i^{(n)})$$

Tabela współczynników wzorów Adamsa – Bashfortha $/\cdot h$

n / k	0	1	2	3
0	1			
1	$\frac{3}{2}$	$-\frac{1}{2}$		
2	$\frac{23}{12}$	$-\frac{10}{12}$	$\frac{5}{12}$	
3	$\frac{55}{24}$	$-\frac{59}{24}$	$\frac{37}{24}$	$-\frac{9}{24}$

Np. dla *n = 2*: $y_{i+1} = y_i + \frac{h}{12} \cdot (23 f_i - 16 f_{i-1} + 5 f_{i-2})$.

- *Metoda Adamsa – Moultona* (metoda zamknięta)

$$y_{i+1} = y_i + h \cdot \sum_{j=n}^{j=0} f_{i-j+1} \cdot L_{i-j+1}^{(n)} = y_i + h \cdot (f_{i-n+1} \cdot L_{i-n+1}^{(n)} + ... + f_i \cdot L_i^{(n)} + f_{i+1} \cdot L_{i+1}^{(n)})$$

Tabela współczynników wzorów Adamsa – Moultona $/\cdot h$

n / k	0	1	2	3
0	1			
1	$\frac{1}{2}$	$\frac{1}{2}$		
2	$\frac{5}{12}$	$\frac{8}{12}$	$-\frac{1}{12}$	
3	$\frac{9}{24}$	$\frac{19}{24}$	$-\frac{5}{24}$	$\frac{1}{24}$

Np. dla *n = 2*: $y_{i+1} = y_i + \frac{h}{12} \cdot (5 f_{i+1} + 8 f_i - f_{i-1})$.

Przykład 15

Znaleźć wartość funkcji $f(1)$, jeżeli

$$f' = f^2 + 2 + x, \quad f(0) = 1, \quad h = 1$$

metodą Rungego - Kutty 4 rzędu.

$$x_0 = 0, \quad f_0 = f(x_0) = f(0) = 1, \quad F(x, f) = f^2 + 2 + x, \quad h = 1$$

$$K_1 = h \cdot F(x_0, f_0) = 1 \cdot F(0,1) = 1 + 2 + 0 = 3$$

$$K_2 = h \cdot F(x_0 + \frac{1}{2}h, f_0 + \frac{1}{2}K_1) = 1 \cdot F(0 + \frac{1}{2} \cdot 1, 1 + \frac{1}{2} \cdot 3) = \left(\frac{5}{2}\right)^2 + 2 + \frac{1}{2} = \frac{35}{4}$$

$$K_3 = h \cdot F(x_0 + \frac{1}{2}h, f_0 + \frac{1}{2}K_2) = 1 \cdot F(0 + \frac{1}{2} \cdot 1, 1 + \frac{1}{2} \cdot \frac{35}{4}) = \left(\frac{43}{8}\right)^2 + 2 + \frac{1}{2} = \frac{2009}{64}$$

$$K_4 = h \cdot F(x_0 + h, f_0 + K_3) = 1 \cdot F(0 + 1, 1 + \frac{2009}{64}) = \left(\frac{2073}{64}\right)^2 + 2 + 1 = \frac{4309617}{64 \cdot 64}$$

$$f_{i+1} = f_i + \frac{1}{6}(K_1 + 2K_2 + 2K_3 + K_4) = 1 + \frac{1}{6}(3 + \frac{35}{4} + \frac{2009}{64} + \frac{4309617}{64 \cdot 64}) = 183.548869$$

Praktyczne stosowanie metod zamkniętych (zazwyczaj wielokrokowych) wiąże się z następującym algorytmem iteracyjnym zwanym zwyczajowo *metodą predyktor – korektor*. Polega ona na znalezieniu kilku pierwszych wartości funkcji metodą jednokrokową wysokiego rzędu (np. *metodą Rungego – Kutty IV rzędu*), a następnie wstępnego określenia (predykcji – stąd nazwa „predyktor") szukanej, następnej z kolei wartości funkcyjnej za pomocą wzoru otwartego wielokrokowego. Wartość ta służy jako punkt startowy dla metody wielokrokowej zamkniętej, która iteracyjnie poprawia (stąd nazwa „korektor") szukaną wartość aż do osiągnięcia wymaganej dokładności.

Dla przykładu rozważmy równanie początkowe rzędu pierwszego

$$\frac{dy}{dx} = f(x, y), \quad x \in (a, b), \quad y(x_0) = y_0; \quad x_0 \in (a, b).$$

Dwie pierwsze wartości funkcyjne znaleziono stosując *metodę Rungego – Kutty rzędu IV*.

$y_0 = y(x_0) \quad \rightarrow$ z warunku początkowego

$$\left.\begin{aligned} y_1 &= y_0 + \Delta y_0 \\ y_2 &= y_1 + \Delta y_1 \end{aligned}\right\} \quad \rightarrow \text{ z } \textit{metody Rungego - Kutty IV rzędu}.$$

Wartość y_3, a dokładniej jej zerowe przybliżenie znaleziono korzystając z *metody Adamsa – Bashfortha rzędu II*.

$$y_3^{(0)} = y_2 + \frac{h}{12} \cdot (23 f_2 - 16 f_1 + 5 f_0), \quad f_i \equiv f(x_i, y_i), \quad i = 0,1,2.$$

Następnie posłużono się odpowiednim schematem zamkniętym (*metoda Adamsa – Moultona rzędu II*) układając w ten sposób procedurę iteracyjną, kontrolowaną przed warunek zbieżności na podstawie znanej dokładności wyniku ε.

$$y_3^{(k+1)} = y_2 + \frac{h}{12}\cdot(5f_3^{(k)} + 8f_2 - f_1), \quad \begin{cases} f_i \equiv f(x_i, y_i), & i = 1,2; \\ f_i^{(k)} = f(x_i, y_i^{(k)}). \end{cases}$$, gdzie dla $k = 0$ wynik $y_3^{(0)}$ pochodzi z poprzedniej metody (z predykatora). Wynik poprawiamy sprawdzając na każdym kroku warunek zbieżności

$$\left\| \frac{y_3^{(k+1)} - y_3^{(k)}}{y_3^{(k+1)}} \right\| \le \varepsilon .$$

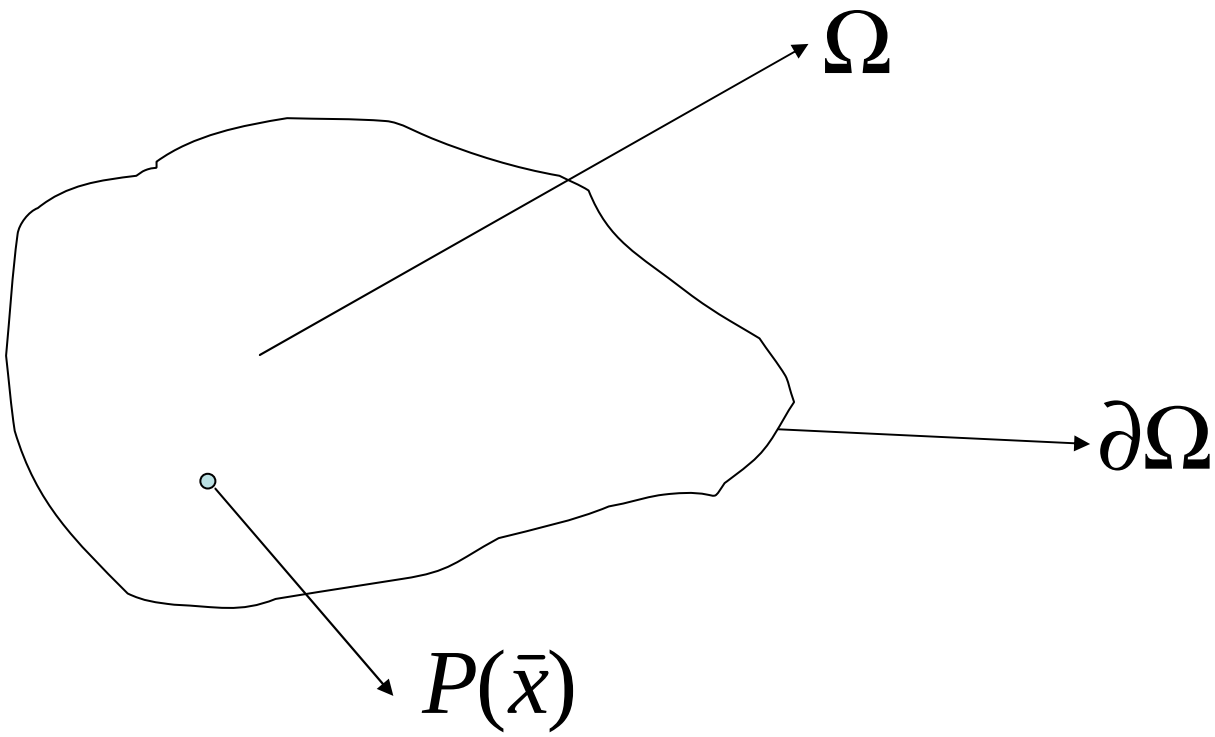
Gdy wynik się ustabilizuje, można przejść do obliczania następnej wartości funkcji y_4 w ten sam sposób, co powyżej.

V. NUMERYCZNE ROZWIĄZYWANIE PROBLEMÓW BRZEGOWYCH

Podstawową różnicą między problemem początkowym i brzegowym jest sposób określeni warunków. W problemie początkowym warunki (początkowe) nałożone były na funkcję niewiadomą i jej kolejne pochodne aż do odpowiedniego rzędu w jednym, wybranym punkcie obszaru. W problemach brzegowych na ogół mamy do czynienia ze zbiorem punktów, w których dane są wartości funkcji lub jej pochodnych. Metody numeryczne do rozwiązywania obydwu problemów diametralnie różnią się od siebie. Problemy początkowe numerycznie prowadziły do znalezienia tablicy wartości funkcji punkt po punkcie zaczynając od punktu z warunkiem początkowym. W metodach dyskretnych do analizy zadań brzegowych otrzymujemy dla zadanego zbioru punktów (węzłów) układ równań, z którego jednocześnie otrzymujemy wartości we wszystkich niewiadomych węzłach.

Niezwykle ważną rzeczą jest sposób sformułowania problemu brzegowego. Ogólnie każdy zapis problemu, w którym występuje nieznana funkcja jest dopuszczalny, ale w zagadnieniach fizyki i mechaniki funkcjonują od lat dwa zasadnicze typy sformułowań brzegowych – lokalne i globalne. Również od sformułowania zależy sposób otrzymania i jakość wyniku różnicowego.

Zagadnienie (problem) brzegowe: dany jest obszar Ω, w którym poszukiwane jest rozwiązanie, układ równań różniczkowych cząstkowych oraz warunki początkowo – brzegowe nałożone na zbiór punktów należących do brzegu $\partial\Omega$ obszaru.



W rozważanym obszarze poszukiwana jest funkcja $u(\bar{x})$ w każdym punkcie $P(\bar{x})$. Można stosować następujące sformułowania zagadnień brzegowych:

- *Sformułowanie lokalne* (*mocne, silne*): szukane jest rozwiązanie układu równań różniczkowych w każdym z punktów obszaru osobno:

  $$\mathcal{L}u = f \quad dla \quad P \in \Omega$$
  $$\mathcal{B}u = g \quad dla \quad P \in \partial\Omega$$

  gdzie $\mathcal{L}$ i $\mathcal{B}$ są operatorami różniczkowymi odpowiednio w obszarze i na jego brzegu. Równanie $\mathcal{B}u = g \quad dla \quad P \in \partial\Omega$ nosi nazwę warunków brzegowych. Jeżeli są one nałożone na funkcję (tzn. $\mathcal{B} \equiv 1$), noszą nazwę podstawowych *warunków brzegowych Dirichleta,* natomiast dowolna kombinacja warunków brzegowych złożona z pochodnych nosi nazwę nosi nazwę naturalnych *warunków brzegowych Neumanna*.
- *Sformułowanie globalne*: może być formułowanie jako problem optymalizacji funkcjonału lub jako zasada wariacyjna.
  - *Minimalizacja funkcjonału*:

    $$I(u) = \frac{1}{2}\mathcal{B}(u,u) - \mathcal{L}(u)$$

    W funkcjonałach energetycznych pierwszy składnik prezentuje energię wewnętrzną układu, podczas gdy drugi jest równy pracy wykonanej przez siły zewnętrzne. Nieznana funkcja $u(P)$ może przedstawiać sobą przemieszczenia u, odkształcenia ε, naprężenia σ lub wszystkie z nich. Funkcja u realizująca ekstremum (minimum, punkt stacjonarny) funkcjonału $\min_{(u)} I(u)$ jest szukana. Można rozważać problem optymalizacji funkcjonału bez ograniczeń (w całej

przestrzeni rozwiązań dopuszczalnych) lub z ograniczeniami (ekstremum jest szukane w podprzestrzeni narzuconych ograniczeń).

- *Zasada wariacyjna*

  $$\mathcal{B}(u, \partial u) = \mathcal{L}(\partial u) \quad dla \quad \partial u \in V$$

  W mechanice powyższe równanie może mieć sens np. zasady prac wirtualnych. Sformułowanie wariacyjne (tzw. słabe) ma podstawowe znaczenie przy konstruowaniu rozwiązań przybliżonych. Można go uzyskać ze sformułowania mocnego w czterech krokach:
  - Przemnożenie równania różniczkowego przez dowolną funkcję (tzw. funkcja testująca),
  - Przecałkowanie wyniku po rozważanym obszarze Ω,
  - Całkowanie przez części z wykorzystaniem *twierdzenia Greena* w celu zredukowania pochodnych do minimalnego rzędu,
  - Wprowadzenie do funkcjonału *warunków brzegowych Neumanna*.

  Sformułowania globalne wymagają dodatkowego całkowania po obszarze. Sformułowanie wariacyjne jest ogólniejsze, gdyż możliwe jest w przypadku wszystkich zagadnień brzegowych, podczas gdy ułożenie funkcjonału możliwe jest tylko dla niektórych zadań mechaniki, np. dla zadań liniowej sprężystości (*funkcjonał Lagrange'a, Hamiltona, Reissnera, Castigliano*, itp.).

- Możliwe są również podejścia mieszane, polegające np. na podziale obszaru Ω na podobszary, gdzie stosuje się różne sformułowania wraz z odpowiednimi warunkami ograniczającymi.

Budowa rozwiązania przybliżonego problemu brzegowego zależy przede wszystkim od wybranej metody dyskretnej. Można wyróżnić dwie główne koncepcje:

- Rozwiązanie dyskretne w postaci kombinacji liniowej współczynników liczbowych oraz funkcji bazowych:

  $$p(x) = a_1\varphi_1(x) + a_2\varphi_2(x) + ... + a_n\varphi_n(x) = \sum_{i=1}^{n} a_i\varphi_i(x).$$

  Funkcje bazowe (najczęściej: wielomiany, funkcje trygonometryczne, funkcje specjalne) muszą być liniowo niezależne, odpowiednio ciągłe oraz muszą spełniać jednorodne warunki brzegowe rozważanego problemu (jednorodne warunki to takie, w których po prawej stronie stoi 0, (np. $u(x_0) = 0$, $u^I(x_0) = 0$). Przy takim zapisie postaci rozwiązania przybliżonego można szukać budując odpowiednie residua, (czyli wyrażenia świadczące o spełnieniu przez rozwiązanie przybliżone wyjściowych równań różniczkowych) odpowiednio w obszarze i na brzegu:

  $$\varepsilon_d(x) = \mathcal{L}p(x) - f, \quad \varepsilon_b(x) = \mathcal{B}p(x) - g$$

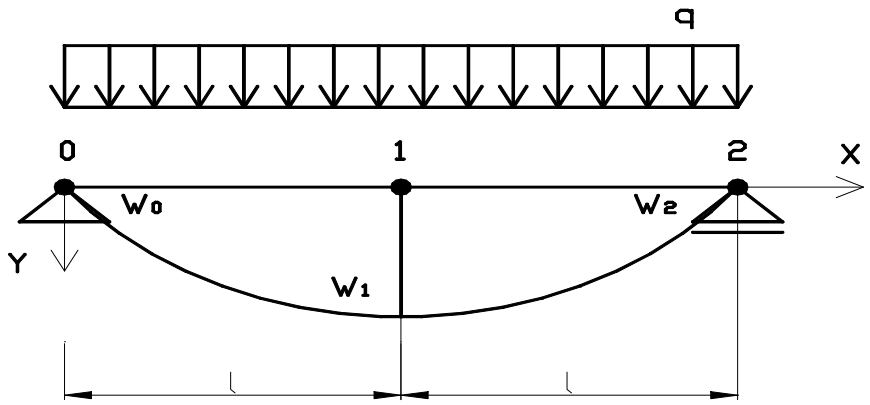
  Funkcjonał ważący powyższe wyrażenia ma postać:

  $$I(p) = \int_{\Omega} \varepsilon_d w_d d\Omega + \int_{\partial\Omega} \varepsilon_b w_b d\partial\Omega.$$

Wagi w_d *i* w_b świadczą o odejściu *p(x)* od wyniku ścisłego odpowiednio w obszarze i na jego brzegi. Dla *metod residuów ważonych* (*metoda Bubnowa - Galerkina, metoda najmniejszych kwadratów, metoda kolokacji*) i *metod energetycznych* (*metoda Rayleigha – Ritza*) zakłada się błąd na brzegu $\varepsilon_b = 0$ (ścisłe spełnienie warunków brzegowych) i rozważa jedynie $I(p) = \int_\Omega \varepsilon_d w_d d\Omega$. Odmienną koncepcję prezentują tzw. *metody Trefza*, w których zakłada się ścisłe spełnienie równania wewnątrz obszaru a rozwiązań przybliżonych poszukuje na jego brzegu.

- Rozwiązanie dyskretne w wybranych punktach obszaru (lub/i jego brzegu) zwanych węzłami. W tej koncepcji niezbędna jest dyskretyzacja obszaru (na węzły, elementy itp.), gdzie zastępuje się wielkości ciągłe wielkościami dyskretnymi. Numeryczne wyniki dyskretne można aproksymować funkcją ciągłą w ramach tzw. *postprocesingu*. Do tych metod należą: *metoda różnic skończonych* (**MRS**, zamiana operatorów różniczkowych na różnicowe, poszukiwanie wartości węzłowych funkcji szukanej, aproksymacja *metodami najmniejszych kwadratów*), *metoda elementów skończonych* (**MES**, podział na elementy i aproksymacja funkcjami kształtu) oraz *metod elementów brzegowych* (**MEB**, podział brzegu na odcinki, obliczanie całek brzegowych).

Przykład 16
Belka swobodnie podparta obciążona obciążeniem ciągłym równomiernie rozłożonym.



Sformułowanie lokalne:

$$\mathcal{L}(x) = \frac{d^2}{dx^2} w(x) = f(x) \qquad f(x) = -\frac{M(x)}{EJ} \quad 0 \le x \le 2l$$

$$M(x) = \frac{1}{2} qx(2l - x) \qquad w(0) = 0 \qquad w(2l) = 0$$

Sformułowanie globalne:

- W postaci funkcjonału:

$$\min_w I(w) \quad \Rightarrow \quad I(w) = \int_0^{2l} [\frac{1}{2}(\frac{dw}{dx})^2 - \frac{M(x)}{EJ} w] dx, \quad w(0) = w(2l) = 0$$

- W postaci zasady wariacyjnej:

$$\int_0^{2l}[\frac{dw^2}{dx^2}+\frac{M(x)}{EJ}]v(x)dx=0,\quad w(0)=w(2l)=0$$

$v(x)$ – funkcja próbna, odpowiednio ciągla, spelnia warunki brzegowe: $v(0)=v(2l)=0$

lub po przecałkowaniu przez części (sformułowanie słabe):

$$\int_0^{2l}[\frac{dw}{dx}\frac{dv}{dx}-\frac{M(x)}{EJ}v(x)]dx=0,\quad w(0)=w(2l)=0,\quad v(0)=v(2l)=0$$

Rozwiązanie przybliżone dla *metod residualnych*:

- Funkcje bazowe: $\varphi_1(x)=x(x-2l),\quad \varphi_2(x)=x^2(x-2l)$,
- Rozwiązanie próbne: $p(x)=a\varphi_1(x)+b\varphi_2(x)=a\,x(x-2l)+b\,x^2(x-2l)$,
- Residuum w obszarze: $\varepsilon(x)=\frac{d^2p(x)}{dx^2}+\frac{M(x)}{EJ}=2a+b(6x-4l)-\frac{1}{2}qx(2l-x)$
- Dla *metody Bubnowa - Galerkina*:

$$\begin{cases}\int_0^{2l}\varepsilon(x)\cdot\varphi_1(x)dx=0\\ \int_0^{2l}\varepsilon(x)\cdot\varphi_2(x)dx=0\end{cases}\Rightarrow\begin{cases}\int_0^{2l}[2a+b(6x-4l)-\frac{1}{2}qx(2l-x)]x(x-2l)dx=0\\ \int_0^{2l}[2a+b(6x-4l)-\frac{1}{2}qx(2l-x)]x^2(x-2l)dx=0\end{cases}\Rightarrow a,b=...$$

- Dla *metody najmniejszych kwadratów*:

$$I(a,b)=\int_0^{2l}\varepsilon(x)\cdot\varepsilon(x)dx\quad\Rightarrow\quad\min_{(a,b)}I(a,b)$$

$$I(a,b)=\int_0^{2l}[2a+b(6x-4l)-\frac{1}{2}qx(2l-x)]^2dx\quad\Rightarrow\quad\begin{cases}\frac{\partial}{\partial a}I(a,b)=0\\ \frac{\partial}{\partial b}I(a,b)=0\end{cases}\Rightarrow a,b=...$$

- Dla *metody kolokacji* (punkty kolokacji: $x_1=\frac{l}{3},\ x_2=\frac{2l}{3}$):

$$\begin{cases}\int_0^{2l}\varepsilon(x)\cdot\delta(x-x_1)dx=0\\ \int_0^{2l}\varepsilon(x)\cdot\delta(x-x_2)dx=0\end{cases}\Rightarrow\begin{cases}\varepsilon(x_1)=0\\ \varepsilon(x_2)=0\end{cases}\Rightarrow\begin{cases}2a+b(6x_1-4l)-\frac{1}{2}qx_1(2l-x_1)=0\\ 2a+b(6x_1-4l)-\frac{1}{2}qx_2(2l-x_2)=0\end{cases}\Rightarrow a,b=...$$

W *metodzie różnic skończonych* **MRS** wprowadzono w ramach dyskretyzacji obszaru 3 węzły (patrz: rysunek). Z trzech wartości węzłowych dwie z nich stanowią warunki brzegowe:

$w_0 = w_2 = 0$, pozostaje do obliczenia wartość w_1. Przy sformułowaniu lokalnym zamianie na operator różnicowy ulega operator różniczkowy na drugą pochodną: $\left.\frac{d^2 w}{dx^2}\right|_{x=l} = w_1^{II} \approx Lw_1 = \frac{w_0 - 2w_1 + w_2}{l^2}$. Równania różnicowe generuje się metodą kolokacji (ścisłe spełnienie równania w węzłach obszaru):

$$Lw_1 = f_1 \quad \Rightarrow \quad \frac{\overbrace{w_0}^{0} - 2w_1 + \overbrace{w_2}^{0}}{l^2} = \frac{1}{2}\frac{ql^2}{EJ} \quad \Rightarrow \quad w_1 = ...$$

W sformułowaniu globalnym można ułożyć funkcjonał energii potencjalnej układu. Po jego dyskretyzacji (całkowanie *kwadraturą Newtona-Cotesa* między węzłami) otrzymuje się:

$$I(w_0, w_1, w_2) = \frac{1}{2}[(\frac{w_1 - w_0}{l})^2 l + (\frac{w_2 - w_1}{l})^2 l - (M_0 w_0 + M_1 w_1)\frac{l}{2EJ} - (M_1 w_1 + M_2 w_2)\frac{l}{2EJ}]$$

Niewiadomą w_1 (oczywiście $w_0 = w_2 = 0$) otrzymuje się minimalizując powyższy funkcjonał względem w_1:

$$\frac{d}{dw_1} I(w_1) = 0 \quad \Rightarrow \quad w_1 = ...$$

Przy sformułowaniu wariacyjnym słabym (funkcja testowa: $v(0) = v(2l) = 0$) od razu otrzymuje się gotowe równanie różnicowe:

$$\frac{w_1 - w_0}{l}\frac{v_1 - v_0}{l} l + \frac{w_2 - w_1}{l}\frac{v_2 - v_1}{l} l - (M_0 w_0 + M_1 w_1)\frac{l}{2EJ} - (M_1 w_1 + M_2 w_2)\frac{l}{2EJ} = 0 .$$

Podstawiając $w_0 = w_2 = 0$ oraz $v_0 = v_2 = 0$ i przyrównując wyrażenie stojące przy dowolnym v_1 do zera otrzymuje się wartość w_1.

Przykład 17

Rozwiązać równanie

$$w'' + w = 1 \qquad w(0) = 0, \qquad w(4) = 1$$

metodą różnic skończonych i analitycznie. Wynik sprawdzić analitycznie dla $x = 2$ (obliczyć normę błędu)

Rozwiązanie analityczne

CO/RJ: $w'' + w = 0 \quad \rightarrow \quad r^2 + 1 = 0 \quad \rightarrow \quad w_0(x) = A\sin(x) + B\cos(x)$ - całka ogólna

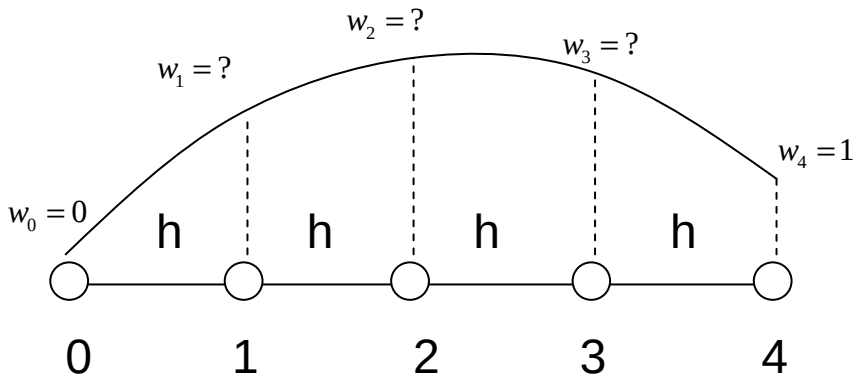
$CS/RNJ \quad w_S(x) = C \quad \rightarrow \quad w_S^{II}(x) + w_S(x) = 1 \quad \rightarrow \quad C = 1 \quad \rightarrow \quad w_S(x) = 1$ - całka szczególna

$$w(x) = w_0(x) + w_S(x) = A\sin(x) + B\cos(x) + 1$$

$$\begin{cases} w(0) = 0 \\ w(4) = 1 \end{cases} \Rightarrow \begin{cases} B + 1 = 0 \\ A\sin(4) + B\cos(4) + 1 = 1 \end{cases} \Rightarrow \begin{cases} A = 0.863691 \\ B = -1 \end{cases}$$

$$\Rightarrow \quad w(x) = 0.863691 \cdot \sin(x) - 1 \cdot \cos(x) + 1$$

Rozwiązanie numeryczne (*metoda różnic skończonych* **MRS**)



Wprowadzono do obszaru zadania $x \in \langle 0,4 \rangle$ pięć równoodległych węzłów ($h = 1$). Warunki brzegowe $w_0 = w(x_0 = 0) = 0, \quad w_4 = w(x_4 = 4) = 1$. Przyjęto klasyczny operator różnicowy na drugą pochodną (zbudowany na trzech węzłach).

$$w_i^{II} \approx \frac{w_{i-1} - 2w_i + w_{i+1}}{h^2}.$$

Generacja równań różnicowych (techniką *kolokacji*)

$$\begin{cases} \dfrac{w_0 - 2w_1 + w_2}{1^2} + w_1 = 1 \\ \dfrac{w_1 - 2w_2 + w_3}{1^2} + w_2 = 1 \\ \dfrac{w_2 - 2w_3 + w_4}{1^2} + w_3 = 1 \\ w_0 = 0, \quad w_4 = 1 \end{cases} \Rightarrow \begin{cases} -2w_1 + w_2 + w_1 = 1 \\ w_1 - 2w_2 + w_3 + w_2 = 1 \\ w_2 - 2w_3 + 1 + w_3 = 1 \end{cases}$$

$$\begin{cases} -w_1 + w_2 = 1 \\ w_1 - w_2 + w_3 = 1 \\ w_2 - w_3 = 0 \end{cases} \Rightarrow \begin{cases} w_1 = 1 \\ w_2 = 2 \\ w_3 = 2 \end{cases}$$

Ścisłe wartości węzłowe (z rozwiązania analitycznego) $w(1) = 1.186469$, $w(2) = 2.201499$, $w(3) = 2.111877$.

Norma błędu wyniku numerycznego dla $x = 2$:

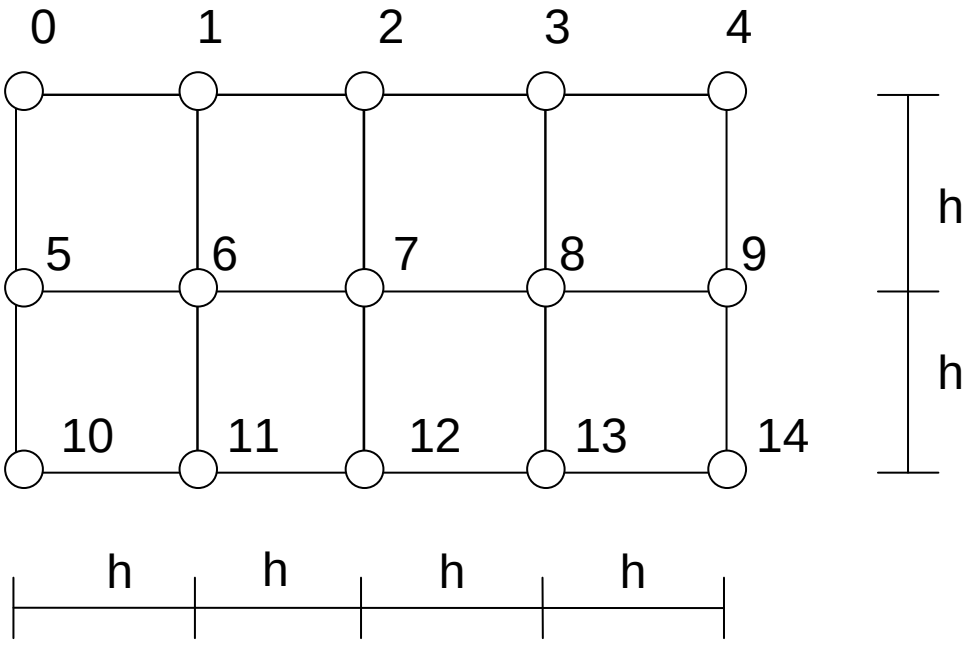
$$\varepsilon = \left| \frac{w(2) - w_2}{w(2)} \right| \cdot 100\% = \left| \frac{2.201499 - 2.0}{2.201499} \right| \cdot 100\% = 9.2\%.$$

Przykład 18

Znaleźć wartości węzłowe dla równania

$$\left(\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}\right) f = 1, \quad h = 1$$

przy zerowych warunkach brzegowych na funkcję.



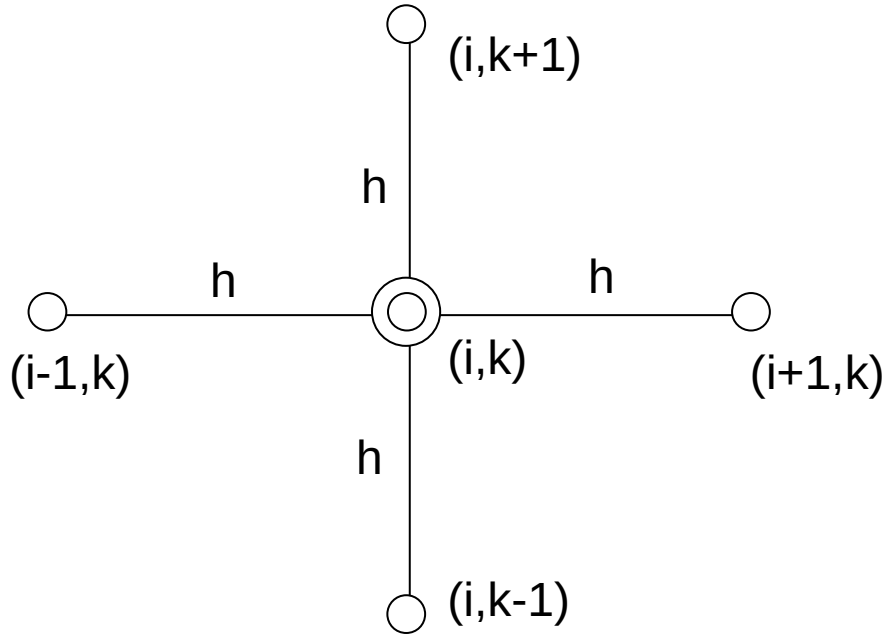
Zadanie brzegowe należy do dziedziny zadań dwuwymiarowych, typu *eliptycznego*. Występujący w sformułowaniu problemu operator różniczkowy zwie się *operatorem Laplace'a*. Mimo to metodologia postępowania jest identyczna jak w zadaniach jednowymiarowych. Obszar zadania podlega dyskretyzacji – wprowadzono 15 węzłów numerowanych od 0 do 14 równomiernie rozłożonych w obszarze (obszarze obydwu kierunkach $h = 1$). 12 z nich do węzły brzegowe, w których z warunków zadania wiadomo, że $f = 0$. Pozostałe węzły zawierają niewiadome węzłowe wartości. W zadaniu można skorzystać z warunków symetrii (symetria wynika z geometrii obszaru, postaci warunków brzegowych i postaci funkcji prawej strony równania różniczkowego w obszarze). Uwzględniając symetrię można zapisać

$$\begin{aligned} &f_0 = f_1 = f_2 = f_3 = f_4 = f_5 = f_9 = f_{10} = f_{11} = f_{12} = f_{13} = f_{14} = 0 \\ &f_8 = f_6 \end{aligned} .$$

Liczba niewiadomych została więc zredukowana do dwóch ($f_6,\ f_7 = ?$).

Kolejnym krokiem analizy numerycznej problemu brzegowego metodą **MRS** jest zastąpienie w węzłach obszaru operatora różniczkowego odpowiednim operatorem różnicowym. Operator różnicowy *Laplace'a* można wygenerować dowolna metodą do budowania schematów różnicowych (np. *metodą współczynników nieoznaczonych* omawianą w **rozdziale**

II). Można również, korzystając z jego prostoty, stworzyć go za pomocą kompozycji odpowiednich składowych operatorów go tworzących. Ostateczne rozwiązanie



to następujący wzór różnicowy

$$\nabla f_{(i,k)} = \left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right) f_{(i,k)} \approx \frac{1}{h^2}\left(f_{(i-1,k)} + f_{(i+1,k)} + f_{(i,k-1)} + f_{(i,k+1)} - 4 f_{(i,k)} \right).$$

Po raz kolejny stosujemy technikę kolokacji do generacji układu równań różnicowych. Przykładamy operator *Laplace'a* w węzłach z niewiadomymi wartościami funkcji – (6) i (7).

$$\begin{cases} \frac{1}{1^2}\left(\overbrace{f_1}^{0} + f_7 + \overbrace{f_{11}}^{0} + \overbrace{f_5}^{0} - 4 f_6 \right) = 1 \\ \frac{1}{1^2}\left(\overbrace{f_2}^{0} + \overbrace{f_8}^{f_6} + \overbrace{f_{12}}^{0} + f_6 - 4 f_7 \right) = 1 \end{cases} \Rightarrow \begin{cases} f_7 - 4 f_6 = 1 \\ 2 f_6 - 4 f_7 = 1 \end{cases} \Rightarrow \begin{cases} f_6 = -\frac{5}{14} \\ f_7 = -\frac{6}{14} \end{cases}.$$