Matematyka stosowana

Numeryczne rozwiązywanie równań różniczkowych

Leszek Marcinkowski
L.Marcinkowski@mimuw.edu.pl
http://www.mimuw.edu.pl/~lmarcin/

Uniwersytet Warszawski, 2011

Streszczenie. W skrypcie przedstawiano rozszerzony materiał z wykładu Numeryczne Równania Różniczkowe o metodach numerycznych (przybliżonych) rozwiązywania równań różniczkowych, zarówno zwyczajnych, jak i typowych równań różniczkowych cząstkowych. W szczególności przedstawimy metody jednokrokowe i wielokrokowe rozwiązywania równań różniczkowych zwyczajnych oraz metodę różnic skończonych i metodę elementu skończonego rozwiązywania równań różniczkowych cząstkowych. Wykład Numeryczne Równania Różniczkowe zawiera tylko część materiału zawartego w skrypcie. Materiał ze skryptu starano się przedstawić w sposób możliwie elementarny, tak więc do zrozumienia wykładu jak i materiału ze skryptu wystarcza wiedza z podstawowych kursowych wykładów z pierwszych dwóch lat studiów, w szczególności nie trzeba znać materiału z wykładu z równań różniczkowych cząstkowych.

Wersja internetowa wykładu:
http://mst.mimuw.edu.pl/lecture.php?lecture=nrr

(może zawierać dodatkowe materiały)

KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI

Projekt współfinansowany przez Unię Europejską w ramach Europejskiego Funduszu Społecznego.

UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY

Skład w systemie LaTeX, z wykorzystaniem m.in. pakietów beamer oraz listings. Szablony podręcznika i prezentacji: Piotr Krzyżanowski; koncept: Robert Dąbrowski.

Spis treści

1. Wprowadzenie

W skrypcie przedstawimy niektóre metody przybliżone, inaczej zwane numerycznymi, rozwiązywania równań różniczkowych zwyczajnych i cząstkowych. Skrypt zawiera rozszerzony materiał z semestralnego wykładu Numeryczne Równania Różniczkowe na wydziale Matematyki, Informatyki i Mechaniki Uniwersytetu Warszawskiego. Na końcu skryptu zamieszczono dodatkowy rozdział 16 z materiałem wykraczającym w większości poza zakres wykładu. Materiał w skrypcie starano się przedstawić w sposób możliwie elementarny, tak więc do zrozumienia wykładu jak i skryptu wystarcza wiedza z podstawowych kursowych wykładów z pierwszych dwóch lat studiów, w szczególności nie trzeba znać materiału z wykładu z równań różniczkowych cząstkowych. Wszystkie potrzebne wiadomości o równaniach różniczkowych zostaną podane w czasie wykładu.

Semestralny wykład powinien koniecznie objąć rozdziały 3-5, 7, 11-12 i 14-15. Ewentualnie można omówić któreś z następujących zagadnień: równania sztywne z rozdziału 6, elementy teorii metody elementu skończonego, tzn. rozdział 13, teorię metody różnic skończonych dla równań eliptycznych, tj. rozdział 8 lub w wersji rozszerzonej także któryś z rozdziałów 9 lub 10. Rozdział 16 zawiera materiał tylko dla osób zainteresowanych teorią metody elementu skończonego.

Z uwagi na to, że rozwiązywanie numeryczne równań różniczkowych to obszerny dział nauki, przedstawiliśmy wybór podstawowych zagadnień różniczkowych, jak i metod ich numerycznego rozwiązywania. Jeśli chodzi o równania cząstkowe rozważamy tylko modelowe zagadnienia liniowe.

Skrypt zawiera również ćwiczenia teoretyczne i laboratoryjne, jak również wyniki prostych eksperymentów komputerowych potwierdzających niektóre wyniki teoretyczne.

Rozdział 2 krótko omawia różne typy równań różniczkowych, których metody rozwiązywania omawiamy w kolejnych rozdziałach.

W rozdziałach 3, 4, 5 i 6 przedstawiono niektóre metody rozwiązywania zagadnień początkowych dla równań różniczkowych zwyczajnych, czyli dwie podstawowe klasy schematów - schematy jednokrokowe i wielokrokowe liniowe wraz z teorią zbieżności tychże schematów. Opisano również schematy dla ważnej klasy zagadnień sztywnych, idee konstrukcji schematów ze zmiennym krokiem dyskretyzacji, oraz idee metody strzałów służącej rozwiązywaniu zagadnień brzegowych. Obszerniejsze informacje dotyczące numerycznego rozwiązywania równań zwyczajnych można znaleźć w pozycjach w języku polskim w [22], [23], [14], [19], a w języku angielskim w monografiach [5], [12] i [13]. Dobrym podręcznikiem w języku angielskim, poświęconym w dużej części numerycznemu rozwiązywaniu równań zwyczajnych jest np. część druga [20], czy rozdział 11 w [25].

W rozdziałach 7, 8, 9 i 10 przedstawiono metodę różnic skończonych (MRS) dla modelowego zagadnienia eliptycznego drugiego rzędu w jednym i dwóch wymiarach, wraz z teorią zbieżności i metodami badania stabilności, zarówno w dyskretnych normach typu maksimum jak i typu L^2. Metoda różnic skończonych rozwiązywania zagadnień różniczkowych cząstkowych jest najprostsza zarówno koncepcyjnie, jak i w praktycznej implementacji. Dotyczy to szczególnie sytuacji, gdy obszar w którym postawione jest zagadnienie różniczkowe posiada prostą geometrię, np.

jest kostką, czy kwadratem. Więcej informacji dotyczących MRS można znaleźć w pozycjach w języku polskim w [10], a w języku angielskim np. w podręcznikach: [20] lub [27].

W kolejnych rozdziałach 11, 12 i 13 zaprezentowano metodę elementu skończonego (MES), ponownie dla modelowego zagadnienia różniczkowego eliptycznego drugiego rzędu w jednym i dwóch wymiarach wraz z elementami teorii zbieżności. Metoda elementu skończonego jest dzisiaj podstawową metodą rozwiązywania równań różniczkowych cząstkowych, z uwagi na uniwersalność, jak i rozwiniętą teorię zbieżności. W dodatkowym rozdziale 16 przedstawiono kilka definicji i własności przestrzeni Sobolewa, jak również niezbędnych do zrozumienia niektórych szczegółów dowodów zawartych w rozdziale 13. W języku polskim sporo informacji o metodzie elementu skończonego można znaleźć w [10], a w języku angielskim np. w obszernych podręcznikach: [2], [26], [16] (reprint oryginału z 1987 roku [15]), czy monografiach np. [4], [6].

W rozdziale 14 omówiono metody konstrukcji schematów rozwiązywania równań parabolicznych i hiperbolicznych drugiego rzędu. Ogólnie rzecz ujmując najpierw wprowadzamy dyskretyzację po zmiennej przestrzennej (np. metodą różnic skończonych lub elementu skończonego) i otrzymujemy układ równań zwyczajnych, którego rozwiązanie przybliża wyjściowe zadanie ewolucyjne, następnie powyższy układ równań zwyczajnych możemy rozwiązać korzystając z jakiejś metody zaprezentowanej w rozdziałach 3,4, 5 tego skryptu. Więcej informacji na ten temat w języku polskim można znaleźć w [10], a w języku angielskim np. w [20], [26] lub [16].

W rozdziale 15 krótko omówiono kilka prostych metod różnicowych rozwiązywania modelowego równania różniczkowego hiperbolicznego pierwszego rzędu. Więcej informacji o numerycznych metodach rozwiązywania równań hiperbolicznych można znaleźć np. w [26], [27], lub w [16].

2. Równania różniczkowe - wprowadzenie

Przy pomocy równań różniczkowych modelowanych jest wiele różnych zagadnień. Równaniami różniczkowymi nazywamy takie równania, w których szukaną niewiadomą jest funkcja lub wektor funkcyjny, których pochodne i same funkcję muszą spełniać odpowiednie równania.

2.1. Równania różniczkowe zwyczajne

Najprostszą klasą równań są równania różniczkowe zwyczajne, (ang. *ordinary differential equation*), czyli równania postaci:

$$F\left(t, u, \frac{du}{dt}, \ldots, \frac{d^k u}{dt^k}\right) = 0 \tag{2.1}$$

na funkcję $u \in C^k((a,b), \mathbb{R}^n)$ dla $F : D \to \mathbb{R}^n$ i D zbioru otwartego w $\mathbb{R}^{1+(k+1)\,n}$. Takie równanie zwyczajne nazywamy równaniem rzędu k.

Przy założeniu, że $\frac{\partial F}{\partial y_k}(\hat{t}, \hat{y}) \neq 0$ dla $(\hat{t}, \hat{y}) = (\hat{t}, \hat{y}_0, \ldots, \hat{y}_k)$, otrzymujemy równanie dające się rozwikłać względem $\frac{d^k u}{dt^k}$, tzn. istnieje funkcja f określona na otoczeniu D_1 punktu $(\hat{t}, \hat{y}_0, \ldots, \hat{y}_{k-1})$ taka, że $F(t, y_0, \ldots, y_{k-1}, f(t, y_0, \ldots, y_{k-1})) = 0$ na D_1. Zatem po rozwikłaniu otrzymujemy nowe równanie:

$$\frac{d^k u}{dt^k} = f\left(t, u, \frac{du}{dt}, \ldots, \frac{d^{k-1}}{dt^{k-1}} u\right),$$

którego rozwiązaniem jest funkcja $u \in C^k((a,b), \mathbb{R}^n)$ i które łatwiej numerycznie rozwiązać. Od tej pory będziemy zakładać, że równanie różniczkowe jest w tej postaci. Więcej informacji na temat metod numerycznych rozwiązywania równań różniczkowych zwyczajnych podanych w sposób niejawny, tzn. w postaci (2.1) (zwanymi też równaniami różniczkowo-algebraicznymi) można znaleźć w [1] lub [3].

Zauważmy, że przez proste podstawienie $x = y_1$ i $x^{(j)} = y_{j+1}$ dla $j = 1, \ldots, k-1$ otrzymujemy nowy układ równań pierwszego rzędu:

$$\begin{aligned} \tfrac{dy_1}{dt} &= y_2 \\ \tfrac{dy_2}{dt} &= y_3 \\ &\vdots \\ \tfrac{dy_k}{dt} &= f_1(t, y_1, \ldots, y_k), \end{aligned} \tag{2.2}$$

który jest szczególnym równaniem pierwszego rzędu postaci:

$$\frac{dx}{dt} = f(t, x), \tag{2.3}$$

gdzie funkcja $f : (a, b) \times G \subset \mathbb{R} \times \mathbb{R}^m \to \mathbb{R}^m$ jest zadaną funkcją ciągłą. Tutaj G jest zbiorem otwartym.

Zagadnieniem początkowym (zagadnieniem Cauchy'ego) nazywamy równanie z warunkiem początkowym:

$$\begin{cases} \frac{dx}{dt} & = & f(t, x) \\ x(t_0) & = & x_0 \end{cases} \tag{2.4}$$

gdzie $t_0 \in (a, b), x_0 \in G$ jest ustalone.

Rozwiązaniem równania (2.3) nazwiemy funkcję ϕ klasy C^1 określoną na podzbiorze otwartym $(c, d) \subset (a, b)$ taką, że

$$\frac{d\phi}{dt}(t) = f(t, \phi(t)) \qquad \forall t \in (c, d).$$

Jeśli dodatkowo $t_0 \in (c, d)$ i $\phi(t_0) = x_0$, czyli ϕ spełnia warunek początkowy to ϕ jest rozwiązaniem zagadnienia początkowego (2.4). W przyszłości często będziemy oznaczać rozwiązanie (2.3) jako $x(t)$.

Podamy teraz kilka prostych przykładów zagadnień fizycznych, czy ogólnie przyrodniczych modelowanych równaniami różniczkowymi zwyczajnymi.

Przykład 2.1. Najprostszy model populacji danego gatunku zwierząt:

$$\begin{aligned} &\tfrac{dN}{dt} = a\, N \qquad t > t_0 \\ &N(t_0) = x_0 > 0 \end{aligned}$$

gdzie $N(t)$ - stan populacji w momencie czasu t i a jest stałą większą od zera, szybkością namnażania się osobników, zależną od gatunku. Tu możemy podać rozwiązania $N(t) = \exp(a\,(t - t_0))$.

Oczywiście ten model jest nierealistyczny, ponieważ populacja - nawet izolowana - nie może rosnąc do nieskończoności. Podajmy więc bardziej skomplikowany model wzrostu logistycznego:

Przykład 2.2. Model logistyczny populacji.

$$\begin{aligned} &\tfrac{dN}{dt} = a\, N\,(1 - N/K) \qquad t > t_0 \\ &N(t_0) = x_0 > 0 \end{aligned}$$

gdzie a, K są stałymi większymi od zera. K oznacza pojemność populacji, czy górną granicę populacji. Tu też możemy podać rozwiązania, ale pozostawimy to jako zadanie.

Przykład 2.3. Rozpad radioaktywnego węgla. Wiemy, że w czasie T połowa atomów węgla rozpada się. Ilość atomów modelowana jest równaniem:

$$\begin{aligned} &\tfrac{dx}{dt} = -a\, x \qquad t > t_0 \\ &x(t_0) = x_0 > 0 \end{aligned} \quad ,$$

gdzie a jest szybkością rozpadu, stałą większą od zera. Rozwiązaniem tego równania jest $x(t) = x_0\, \exp(-a\,(t - t_0))$.

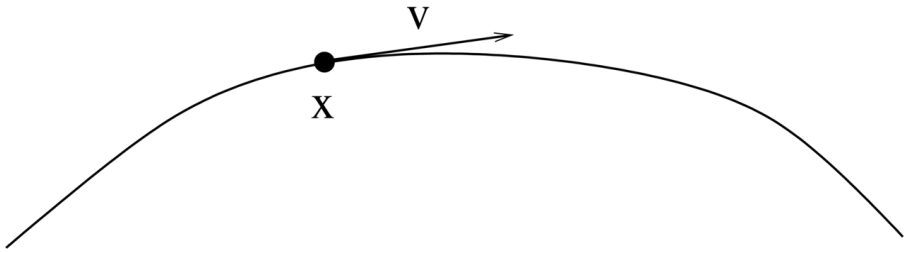
Przykład 2.4. Równanie Newtona.

Rozpatrzmy ruch cząsteczki w przestrzeni. Oznaczmy wektory:

— $x(t) \in \mathbb{R}^3$ położenie cząsteczki w przestrzeni w czasie t,

— $v = \frac{d^2x}{dt^2}$ prędkość cząsteczki,

— $a = \frac{dv}{dt} = \frac{d^2x}{dt^2}$ pochodna prędkości, czyli druga pochodna położenia, tj. przyspieszenie.



Rysunek 2.1. Ruch cząsteczki.

Jeśli ruch cząsteczki sterowany jest jakąś zewnętrzną siłą

$$F : D \subset \mathbb{R}^3 \to \mathbb{R}^3$$

to - zgodnie z prawem dynamiki Newtona - zachodzi następujący związek:

$$m\,a = F(x(t)),$$

gdzie m jest masą cząsteczki. W ten sposób otrzymaliśmy równanie różniczkowe zwane równaniem Newtona:

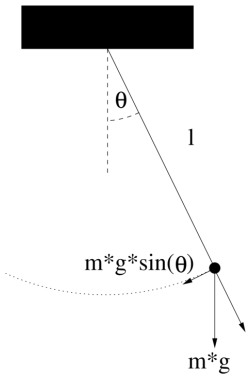
$$\frac{d^2x}{dt^2} = \frac{F(x)}{m}$$

Jeśli dodatkowo znamy położenie i prędkości cząsteczki, tzn. $x(t_0)$ i $v(t_0) = \frac{dx}{dt}(t_0)$ w danym momencie czasu, to możemy wyznaczyć jej położenie po jakimś czasie.

W najprostszym przypadku załóżmy, że działa siła grawitacji skierowana w dół, czyli wzdłuż osi OX_3 (jest to duże uproszczenie, ale dość dobrze modeluje ruch): tzn. siła stała $F(x) = (0, 0, -m\,g)^T$. Otrzymujemy wówczas równanie

$$\begin{aligned} \tfrac{d^2x_1}{dt^2} &= 0 \\ \tfrac{d^2x_2}{dt^2} &= 0 \\ \tfrac{d^2x_3}{dt^2} &= -m\,g. \end{aligned}$$

Znając położenie i prędkość w chwili $t = 0$ łatwo je rozwiązać: $x_1(t) = x_1(0) + v_1(0)\,t$, $x_2(t) = x_2(0) + v_2(0)\,t$ i $x_3(t) = x_3(0) + v_3(0)\,t - 0.5\,m\,g\,t^2$.



Rysunek 2.2. Wahadło.

Przykład 2.5. Równanie wahadła.

Wyprowadzamy równanie zgodnie z Rysunkiem 2.2. Ruch powoduje siła $F(\theta) = -\sin(\theta)\, m\, g$, gdzie m jest masą, g to przyspieszenie ziemskie, a θ jest kątem wychylenia się wahadła. Długość łuku:

$$s = l\,\theta$$

gdzie l to długość wahadła, stąd

$$m\,a = m\,\frac{d^2 s}{dt^2} = m\,\frac{d^2\theta}{dt^2}\,l = -\sin(\theta)\, m\, g$$

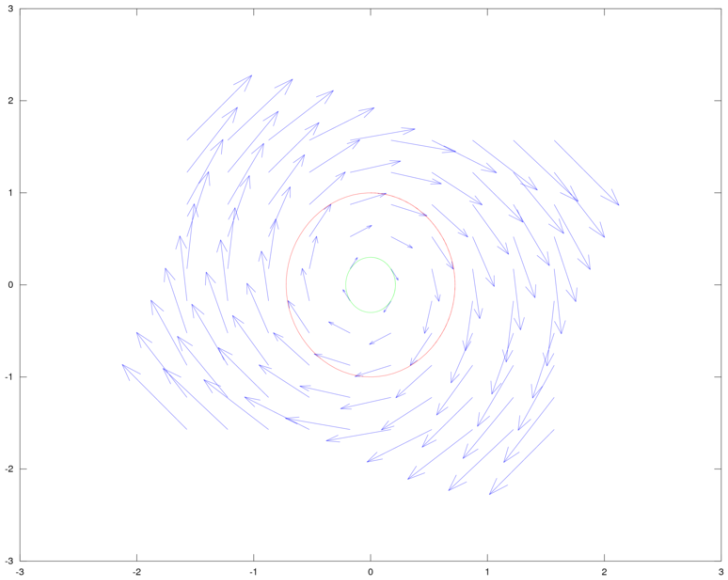
zatem otrzymujemy równanie:

$$\frac{d^2\theta}{dt^2} = -\sin(\theta)\, g/l.$$

Sprowadzając je do równania pierwszego rzędu otrzymujemy:

$$\frac{d}{dt}\begin{pmatrix}\theta\\ \nu\end{pmatrix} = \begin{pmatrix}\frac{d\theta}{dt}\\ \frac{d\nu}{dt}\end{pmatrix} = \begin{pmatrix}\nu\\ -\sin(\theta)\, g/l\end{pmatrix} = f\left(\begin{pmatrix}\theta\\ \nu\end{pmatrix}\right)$$

Możemy naszkicować pole wektorowe tego równania. Tzn. ogólnie jakakolwiek trajektoria rozwiązania $\{(\theta(t), \nu(t))\}$ jest styczna do pola wektorowego zadanego przez prawą stronę równania $f((\theta, \nu)^T)$, czyli w naszym przypadku pole wektorowe w punkcie (θ, ν) przyjmuje wartość $(\nu, -\sin(\theta)\, g/l)^T$, por. Rysunek 2.3.



Rysunek 2.3. Pole wektorowe równania wahadła.

2.2. Równania różniczkowe cząstkowe

Ogólnie mówiąc, równania różniczkowe cząstkowe to równania, których rozwiązania są funkcjami wielu zmiennych, i w których pojawiają się pochodne cząstkowe. Przy niektórych typach

równań wyróżnia się jedną ze zmiennych i oznacza jako czas t; o takich równaniach mówimy często jako o równaniach ewolucyjnych.

W tym rozdziale wymienimy podstawowe typy równań różniczkowych cząstkowych, które pojawią się w treści tego skryptu.

Po więcej informacji na temat podstawowych idei i pojęć dotyczących dziedziny matematyki zwanej równaniami różniczkowymi cząstkowymi odsyłamy do obszernego podręcznika Lawrence'a Evansa [11].

2.2.1. Równania eliptyczne

W przypadku równań eliptycznych nie mamy wyróżnionej zmiennej, ponieważ opisują one często stany stacjonarne zjawisk fizycznych.

Podstawowym przykładem równania eliptycznego jest

równanie Laplace'a:

$$-\triangle u(x) = f(x) \qquad x \in \Omega \subset \mathbb{R}^n,$$

gdzie $\triangle = \sum_{k=1}^{n} \frac{\partial^2}{\partial x_k^2}$ i Ω jest obszarem.

Jeśli dołożymy warunek brzegowy, to otrzymamy klasyczne równanie Poissona. Szukamy tu $u \in C^2(\Omega) \cap C(\overline{\Omega})$ takiego, że

$$\begin{cases} -\triangle u(x) = f(x) & x \in \Omega \\ u(s) = g(s) & s \in \partial\Omega \end{cases} \tag{2.5}$$

Zagadnienie z laplasjanem może mieć też inne warunki brzegowe.

To jest podstawowy przykład zagadnienia eliptycznego, zwanego też zagadnieniem stacjonarnym, czy zagadnieniem brzegowym. W szczególności równanie Laplace'a modeluje rozkład potencjału elektrycznego w $\mathbb{R}^3$.

Zachodzi prawo fizyczne Gaussa:

$$\mathrm{div} E = \rho/\epsilon_0,$$

gdzie div $u = \sum_{k=1}^{3} \frac{\partial u_k}{\partial x_k}$ - to operator dywergencji (rozbieżności) pola, E - to natężenie pola elektrycznego, ρ_0 - to gęstość ładunku elektrycznego, ϵ_0 - to przenikalność elektryczna.

Minus gradient potencjału V daje natężenie pola elektrycznego, tzn.

$$E = -\nabla V$$

z tego wynika, że otrzymujemy

$$\mathrm{div} E = \mathrm{div}(-\nabla V) = -\triangle V.$$

Jeśli ładunek równy zero, to otrzymujemy równanie Laplace'a:

$$\triangle V = 0.$$

Podamy teraz ogólniejszą definicję równania (operatora) eliptycznego drugiego rzędu. Rozważmy równanie różniczkowe liniowe drugiego rzędu dla ogólnego operatora liniowego drugiego rzędu L, określonego dla $u \in C^2(G)$ dla $G \subset \mathbb{R}^n$:

$$Lu = -\sum_{k,l=1}^{n} a_{kl}(x) \frac{\partial^2 u}{\partial x_k \partial x_l}(x) + \sum_{k=1}^{n} b_k(x) \frac{\partial u}{\partial x_k}(x) + c(x)u(x) = f(x) \tag{2.6}$$

gdzie a_{kl}, b_k, c, f są danymi funkcjami (zazwyczaj ciągłymi) określonymi na obszarze $G \subset \mathbb{R}^n$.

Definicja 2.1. Równanie (2.6) (operator L) jest eliptyczne w punkcie x, gdy macierz $A(x) = (a_{kl}(x))_{kl=1,\ldots,n}$ jest dodatnio określona: tzn.:

$$\xi^t A(x)\xi > 0 \qquad \forall \xi \in \mathbb{R}^n$$

Operator L jest eliptyczny w obszarze Ω jeśli L jest eliptyczny w każdym punkcie obszaru Ω.

Warto wspomnieć, że w praktyce pojawiają się także równania eliptyczne czwartego rzędu, np. równanie bi-harmoniczne, które modeluje np. wygiętą cienką membranę (czy płytkę) poprzez zewnętrzną siłę:

$$\triangle^2 u = f \qquad \text{w} \qquad \Omega \subset \mathbb{R}^2,$$

gdzie $\triangle^2 = \triangle\triangle$ - to operator bi-harmoniczny, u - to odchylenie membrany od położenia zero, f - to siła wyginająca membranę pionowo do góry. Tutaj też mogą zachodzić warunki brzegowe różnego typu:

$$u = g_1 \quad \partial_n u = g_2 \qquad \text{na} \qquad \partial\Omega$$

dla płytki przygiętej (tutaj n - to wektor normalny zewnętrzny do brzegu Ω), czy

$$u = g \qquad \text{na} \qquad \partial\Omega$$

dla zadania podpartej płytki.

2.2.2. Równania hiperboliczne pierwszego rzędu

Ogólnie za równanie różniczkowe hiperboliczne pierwszego rzędu uważamy równanie postaci:

$$F\left(x, u, \frac{\partial u}{\partial x_1}, \ldots, \frac{\partial u}{\partial x_N}\right) = 0 \qquad x \in \Omega \subset \mathbb{R}^N$$

dla funkcji $F : \Omega \times G \subset \Omega \times \mathbb{R} \times \mathbb{R}^N \to \mathbb{R}$ i obszaru $\Omega \subset \mathbb{R}^N$.

Dodatkowo dodaje się warunek brzegowy na brzegu lub części brzegu Ω np.:

$$u = g,$$

gdzie g - to dana funkcja.

Będą nas w szczególności interesować równania liniowe:

$$F(x, u, \nabla u) = \vec{a}(x)^T \nabla u + b(x)u + c(x) \tag{2.7}$$

dla danych funkcji $a_k, b, c : \Omega \to \mathbb{R}$.

Ważnym przykładem jest równanie:

$$\frac{\partial u}{\partial t} + a\frac{\partial u}{\partial x} = 0 \qquad t \in \mathbb{R} \quad x \in \mathbb{R}$$

gdzie a - to stała, dla którego znamy rozwiązanie:

$$u(t, x) = F(at - x)$$

dla dowolnej funkcji różniczkowalnej w sposób ciągły F.

Dodając warunek początkowy

$$u(0, x) = \phi(x)$$

dla $\phi \in C^1(\mathbb{R})$ otrzymujemy jednoznaczne rozwiązanie

$$u(t, x) = \phi(at - x).$$

2.2.3. Równania hiperboliczne drugiego rzędu

Ogólnie równaniem liniowym hiperbolicznym drugiego rzędu nazwiemy równanie:

$$\frac{\partial^2 u}{\partial t^2} - Lu = f \qquad t > 0 \quad x \in \Omega \tag{2.8}$$

dla operatora L eliptycznego w $\Omega \subset \mathbb{R}^N$. Tutaj $u_{tt} = \frac{\partial^2 u}{\partial t^2}$.

Klasycznym przykładem takiego równania jest równanie falowe:

$$\frac{\partial^2 u}{\partial t^2} - \triangle u = f \qquad x \in \Omega \subset \mathbb{R}^N \quad N = 1, 2, 3.$$

Dla prawej strony równej zero, tj. $f = 0$, nazywamy je jednorodnym równaniem falowym, a w przeciwnym przypadku nazywamy je niejednorodnym równaniem falowym.

Odpowiada ono drganiu struny ($N = 1$), membrany ($N = 2$) i elastycznej bryły ($N = 3$). Wartości $u(t, x)$ odpowiadają położeniu np. struny w momencie czasu t, jako że zmienna t odpowiada czasowi - jest to równanie ewolucyjne.

Aby zadanie posiadało jednoznaczne rozwiązanie należy:

— Podać warunki brzegowe np. typu Dirichleta

$$u(t, s) = g(t, s) \qquad s \in \partial\Omega$$

dla danej funkcji $g : [0, T] \times \partial\Omega \to \mathbb{R}$. Zakładamy, że na brzegu znamy położenie struny. Gdyby $g(t, s) = g(s)$, to struna czy membrana byłaby zaczepiona.

— Podać warunki początkowe:

$$u(0, x) = \phi(t)$$
$$\frac{\partial u}{\partial t}(0, x) = \psi(t)$$

dla danych funkcji określonych na Ω. Warunki początkowe oznaczają, że znamy położenie i prędkości np. struny w momencie startowym $t = 0$.

2.2.4. Równania paraboliczne

Równaniem liniowym parabolicznym drugiego rzędu nazywamy równanie:

$$\frac{\partial u}{\partial t} - Lu = f \qquad t > 0 \quad x \in \Omega, \tag{2.9}$$

gdzie L operator eliptyczny w $\Omega \subset \mathbb{R}^N$.

Klasycznym równaniem parabolicznym jest równanie przewodnictwa ciepła:

$$\frac{\partial u}{\partial t} - \triangle u = f \qquad t > 0, \quad x \in \Omega \subset \mathbb{R}^N \quad N = 1, 2, 3$$

opisujące rozchodzenie się ciepła w pręcie ($N = 1$), cienkiej płytce ($N = 2$), czy bryle ($N = 3$). Wartości $u(t, x)$ odpowiadają temperaturze w punkcie x w momencie czasu t. Jest to równanie ewolucyjne. Aby zadanie było dobrze postawione należy dodać warunek początkowy $u(0, x) = \phi(x)$ w Ω oraz warunki brzegowe np. typu Dirichleta

$$u(t, s) = g(s) \qquad s \in \partial\Omega$$

dla danej funkcji $g : [0, T] \to \partial\Omega$ co oznacza, że znamy temperaturę na brzegu i temperaturę początkową:

$$u(0, x) = \phi(x)$$

dla danej funkcji ϕ określonej na Ω.

Możemy też na brzegu Ω postawić inne warunki brzegowe np. z pochodną, które odpowiadają temu, że znamy strumień energii wpływającej do płytki, czyli

$$\partial_n u(t,s) = h(s) \qquad s \in \partial\Omega.$$

W jednym wymiarze, tzn. dla $\Omega = (0, L)$ i dla równania ze współczynnikiem stałym $a > 0$ i $f = 0$, warunkami brzegowymi $u(0) = u(L) = 0$ i warunkiem początkowym $u_0 = \sin(k\,t\,\pi/L)$ tzn.:

$$\begin{aligned} \frac{\partial u}{\partial t} &= a\,\frac{\partial^2 u}{\partial x^2} \qquad \text{w} \qquad (0,T)\times(0,L) \\ u(0) &= u(L) = 0, \\ u(x,0) &= \sin(\pi\,x/L) \quad x \in (0,L) \end{aligned}$$

znamy rozwiązanie: $u(t,x) = \exp(-a\,(\pi/L)^2\,t)\sin(\pi\,x/L)$, czyli rozwiązanie gaśnie wraz z upływem czasu.

2.3. Zadania

Ćwiczenie 2.1. Rozpatrzmy zadanie początkowe autonomiczne (tzn. prawa strona równania nie zależy od t):

$$\begin{aligned} \frac{dy}{dt} &= g(x,y) \\ \frac{dx}{dt} &= f(x,y) \\ x(t_0) &= x_0 \qquad y(t_0) = y_0 \end{aligned}$$

dla $f, g \in C^1(G)$, G - to obszar, i $|f(x_0, y_0)| > 0$ dla pewnego $(x_0, y_0)^T \in G$. Pokaż, że istnieje otoczenie U_{x_0} punktu x_0 takie, że na tym otoczeniu równanie

$$dy/dx = f(x,y)/g(x,y) \quad y(x_0) = y_0$$

ma rozwiązania $\psi(x)$ takie, że krzywa całkowa tego równania, tzn. zbiór $\{(x, \psi(x) : x \in U_{x_0}\}$ zawarta jest w trajektorii wyjściowego równania, tzn. w zbiorze $\{(x(t), y(t))\}$ dla x, y rozwiązań wyjściowego równania.

Rozwiązanie. Z tego, że $\frac{dx}{dt}(t_0) = f(x_0, y_0) \neq 0$ i z twierdzenia o funkcji odwrotnej wynika, że istnieje otoczenie U_{x_0}, na którym określona jest funkcja $t(x)$ odwrotna do $x(t)$, której pochodna równa się $dt/dx = 1/(dx/dt) = 1/f$. Wtedy szukaną funkcją jest złożeniem $y(t)$ i $t(x)$, czyli $\psi(x) := y(t(x))$ i zawieranie się krzywej całkowej w trajektorii jest oczywiste.

Ćwiczenie 2.2. Wyprowadź równania ruchu wahadła w postaci:

$$\begin{aligned} \frac{d^2x}{dt^2} &= f(x,y) \\ \frac{d^2y}{dt^2} &= g(x,y). \end{aligned}$$

dla (x, y) położenia wahadła (przyjmujemy, że dla $\theta = 0$ zachodzi $x = y = 0$).

Narysuj powyższe pole wektorowe wahadła w Octavie (funkcja **quiver()**).

Wskazówka. Trzeba dokonać rozkładu na odpowiednie składowe jedynej siły, która powoduje ruch wahadła czyli $-mg\sin(\theta)$ stycznej do toru ruchu. Następnie skorzystać z tego jak wyraża się położenie punktu w terminach θ.
Rozwiązanie. Zauważmy, że $(x,y)^T = (\sin(\theta),\cos(\theta))^T$ i siła działająca poziomo jest równa $-m\,g\,\sin(\theta)\,\cos(\theta) = -m\,g\,x\,y$ a działająca pionowo: $-m\,g\,\sin(\theta)\,\sin(\theta) = -m\,g\,x^2$.

Ćwiczenie 2.3 (Metoda Fouriera)**.** Rozważmy równanie paraboliczne jednowymiarowe:

$$\frac{\partial u}{\partial t}(t,x) = \frac{\partial^2 u}{\partial x^2}(t,x) \quad \text{w} \quad (0,T)\times(0,1)$$

z warunkami brzegowymi $u(t,0) = u(t,1) = 0$ i początkowym $u(0,x) = u_0(x)$. Załóżmy, że szukamy rozwiązania postaci:

$$u(t,x) = f(x)g(t).$$

Wstaw u takiej postaci do powyższego równania i pokaż, że dostajemy dwa niezależne równania różniczkowe zwyczajne na f i g. Rozwiąż te równania tzn. znajdź rozwiązania uogólnione i sprawdź dla jakich u_0 możemy wyznaczyć rozwiązanie wyjściowego problemu.

Ćwiczenie 2.4 (Metoda Fouriera)**.** Rozważmy równanie hiperboliczne jednowymiarowe:

$$\frac{\partial^2 u}{\partial t^2}(t,x) = \frac{\partial^2 u}{\partial x^2}(t,x) \quad \text{w} \quad (0,T)\times(0,1)$$

z warunkami brzegowymi $u(0,t) = u(1,t) = 0$ i początkowymi $u(x,0) = u_0(x)$ i $\frac{\partial u}{\partial t}(x,0) = v_0(x)$. Załóżmy, że szukamy rozwiązania postaci:

$$u(t,x) = f(x)g(t).$$

Wstaw u takiej postaci do powyższego równania i pokaż, że dostajemy dwa niezależne równania różniczkowe zwyczajne na f i g. Rozwiąż te równania, tzn. znajdź rozwiązania uogólnione, czyli rodzinę rozwiązań zależną od stałej, i sprawdź dla jakich u_0, v_0 możemy wyznaczyć rozwiązanie wyjściowego problemu.

Ćwiczenie 2.5 (Metoda charakterystyk)**.** Rozpatrzmy równanie różniczkowe pierwszego rzędu $F(x,u,\nabla u) = 0$ dla funkcji $F : G\times\mathbb{R}\times\mathbb{R}^m \to \mathbb{R}$ i $G\subset\mathbb{R}^m$. Przyjmijmy, że szukamy krzywych $\vec{x} : (a,b)\to\mathbb{R}^m$ na których można wyznaczyć rozwiązanie. Przyjmijmy oznaczenia

$$\begin{aligned} w(s) &= u(x(s)) \\ z_j(s) &= \frac{\partial u}{\partial x_j}(x(s)) \qquad j = 1,\ldots,m. \end{aligned}$$

Różniczkując ostatnie równanie otrzymujemy:

$$\frac{dz_j}{ds} = \sum_{i=1}^{m} \frac{\partial^2 u}{\partial x_j \partial x_i}(x(s))\frac{dx_i}{ds} \tag{2.10}$$

a różniczkując po x_j wyjściowe równanie widzimy, że

$$\frac{\partial F}{\partial x_j}(x,u,\nabla u) + \frac{\partial F}{\partial w}(x,u,\nabla u)\frac{\partial u}{\partial x_j} + \sum_{i=1}^{m}\frac{\partial F}{\partial z_i}(x,u,\nabla u)\frac{\partial^2 u}{\partial x_j\partial x_i} = 0 \tag{2.11}$$

Treścią zadania jest wykazanie, że definiując krzywą $x(s)$ jako krzywą spełniającą równanie:

$$\frac{dx_i}{ds} = \frac{\partial F}{\partial z_i}(\vec{x},w,\vec{z}) \qquad i = 1,\ldots,m, \tag{2.12}$$

i korzystając z powyższych równań otrzymujemy, że $\vec{x}, w, \vec{z}$ spełniają następujący układ równań zwyczajnych:

$$\begin{aligned}
\frac{dx_j}{ds} &= \frac{\partial F}{\partial z_j}(\vec{x}, w, \vec{z}) \qquad j = 1, \ldots, m \\
\frac{dw}{ds} &= \sum_{i=1}^{m} z_i \frac{\partial F}{\partial z_i}(\vec{x}, w, \vec{z}) \\
\frac{dz_j}{ds} &= -\frac{\partial F}{\partial x_j}(\vec{x}, w, \vec{z}) - z_j \frac{\partial F}{\partial w}(\vec{x}, w, \vec{z}) \qquad j = 1, \ldots, m.
\end{aligned}$$

Równania te nazywamy równaniami charakterystyk dla wyjściowego równania pierwszego rzędu, a krzywe $\vec{x}$ - charakterystykami tego równania.
Wskazówka. Drugie równanie na pochodną, tzn. w, uzyskujemy różniczkując po zmiennej s równanie $w(s) = u(x(s))$, a ostatnie równanie otrzymujemy eliminując człon z drugimi pochodnymi u z (2.10) korzystając z (2.11).

Ćwiczenie 2.6. Wyprowadź równania charakterystyk dla równań liniowych pierwszego rzędu (2.7) jednorodnych tzn. z $c(x) = 0$. Oblicz rozwiązania dla równania liniowego w dwóch wymiarach dla $\vec{a}(x) = (1, a_2)$ i a_2 stałej, $b = c = 0$ i warunku brzegowego $u(0, x) = \sin(x)$ dla $x \in \mathbb{R}$.

3. Metody dla równań różniczkowych zwyczajnych

W tym i kilku następnych rozdziałach zajmiemy się schematami rozwiązywania równań różniczkowych zwyczajnych. Ten rozdział jest poświęcony wprowadzeniu najprostszych metod (schematów) rozwiązywania równań różniczkowych zwyczajnych.

3.1. Wprowadzenie

Załóżmy, że rozpatrujemy zagadnienie początkowe pierwszego rzędu (zagadnienie Cauchy'ego) :

$$\begin{cases} \frac{dx}{dt} & = & f(t,x) \\ x(t_0) & = & x_0 \end{cases} \tag{3.1}$$

gdzie $G \subset \mathbb{R}^m$, $f : (a,b) \times G \to \mathbb{R}^m$ jest funkcją ciągłą, a $t_0 \in (a,b), x_0 \in G$ jest ustalone.

Z ogólnej teorii równań różniczkowych, por. [23] wiadomo, że

Twierdzenie 3.1 (Peano)**.** *Jeśli f jest funkcją ciągłą na otoczeniu (t_0, x_0), to istnieje rozwiązanie (3.1) określone na pewnym otoczeniu t_0.*

Jeśli dodatkowo założymy, że f jest funkcją lipschitzowska na otoczeniu (t_0, x_0) względem zmiennej x, to możemy pokazać jednoznaczności rozwiązania, tzn. zachodzi twierdzenie:

Twierdzenie 3.2 (Picarda-Lindelöfa)**.** *Jeśli f jest funkcją ciągłą na otoczeniu (t_0, x_0) oraz f jest funkcją lipschizowską względem x w pewnej kuli $B((t_0, x_0), \delta)$, tzn.*

$$\exists L \geqslant 0 \quad \forall (t,x),(t,y) \in B((t_0,x_0),\delta) \qquad \|f(t,x) - f(t,y)\| \leqslant L\|x-y\|,$$

to istnieje $c > 0$ i $x \in C^1((t_0 - c, t_0 + c), R^n)$ takie, że x jest jednoznacznym rozwiązaniem (3.1).

Od tej pory będziemy przyjmować, że funkcja f, zwana też polem wektorowym, spełnia założenia twierdzenia Picarda-Lindelöfa, tzn. Twierdzenia 3.2, czyli że istnieje jednoznaczne rozwiązanie zadania Cauchy'ego na odcinku $[t_0, T]$.

Zauważmy, że każde rozwiązanie jest krzywą styczną do pola wektorowego.

3.2. Równania liniowe ze stałymi współczynnikami

W tym podrozdziale krótko przypomnimy teorie dla ważnej klasy równań różniczkowych zwyczajnych, tzn. jednorodnych równań liniowych ze stałymi współczynnikami, czyli równań postaci:

$$\frac{dx}{dt} = Ax$$

gdzie A - to stała macierz $n \times n$, dla których znamy rozwiązania zadania Cauchy'ego:

$$\frac{dx}{dt} = Ax \qquad x(t_0) = x_0 \tag{3.2}$$

Znamy wzór na rozwiązanie tego zadania:

$$x(t) = e^{A\,(t-t_0)}\,x_0,$$

gdzie eksponent od macierzy zdefiniowany jest wzorem

$$\exp(B) = e^B = \sum_{k=0}^{\infty} \frac{B^k}{k!}.$$

Skorzystamy ze znajomości postaci rozwiązania tej klasy równań w rozdziale 6.

W zależności od postaci Jordana macierzy A można wypisać postać $\exp(A\,t)$, w szczególności jeśli macierz A jest diagonalizowalna, tzn. istnieje baza wektorów własnych, które zapisane jako kolumny macierzy V dają:

$$V \Lambda V^{-1} = A$$

gdzie Λ - to macierz diagonalna z wartościami własnymi macierzy A na diagonali:

$$\Lambda = \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix}.$$

Wtedy wiadomo, że

$$e^{A\,t} = V \begin{pmatrix} e^{\lambda_1 t} & & \\ & \ddots & \\ & & e^{\lambda_n t} \end{pmatrix} V^{-1}.$$

W szczególności jeśli $Re\,\lambda_k < 0$ dla wszystkich $k = 1, \ldots, n$ to każde rozwiązanie spełnia

$$\lim_{t \to +\infty} \|x(t)\| = 0,$$

a z kolei jeśli istnieje λ_k takie, że $Re\lambda_k > 0$, to istnieje rozwiązanie zagadnienia Cauchy'go z niezerowym warunkiem brzegowym, dla którego

$$\lim_{t \to +\infty} \|x(t)\| = +\infty.$$

3.3. Kilka prostych schematów

Załóżmy, że rozpatrujemy zadanie skalarne tzn. $m = 1$. Chcemy w jakiś sposób przybliżyć rozwiązanie równania (3.1). Przybliżamy pochodną poprzez iloraz różnicowy dla pewnego parametru $h > 0$:

$$\frac{x(t+h) - x(t)}{h} \approx \frac{dx}{dt}$$

i otrzymujemy otwarty schemat Eulera:

$$\frac{x_h(t+h) - x_h(t)}{h} = f(t, x_h(t))$$

czy inaczej:

$$x_h(t+h) = x_h(t) + h\, f(t, x_h(t))$$

znając rozwiązanie w punkcie t_0 $x_h(t_0) = x_0$ możemy wyznaczyć przybliżone rozwiązanie x_h w kolejnych punktach $t_n = t_0 + n\,h$ z powyższego wzoru. Ale można też przybliżyć pochodną biorąc parametr $-h$ *w tył*:

$$\frac{x_h(t) - x_h(t-h)}{h} \approx \frac{dx}{dt}$$

i wtedy zastępując pochodną przez taki iloraz otrzymujemy zamknięty schemat Eulera:

$$\frac{x_h(t) - x_h(t-h)}{h} = f(t, x_h(t))$$

czy inaczej:

$$x_h(t+h) = x_h(t) + h\, f(t, x_h(t+h)).$$

Proszę zauważyć, że jeśli znamy rozwiązanie w punkcie t_0, tzn. $x_h(t_0) = x_0$, to aby wyznaczyć kolejne przybliżenia rozwiązania w punktach $t = t_n = t_0 + nh$ należy rozwiązać równania postaci:

$$g(y) := y - h\, f(t, y) - x_h(t) = 0 \tag{3.3}$$

względem y, co sprawia, że zamknięty schemat Eulera może wydać się mało praktyczny w porównaniu z otwartym schematem Eulera. Dla niektórych równań jest to pozorne. Zauważmy tylko, że im h mniejsze, tym potencjalnie równanie (3.3) jest łatwiejsze do rozwiązania (dlaczego?). Przyjrzymy się temu problemowi dokładniej w kolejnych rozdziałach.

W dalszej części wykładu założymy, że chcemy przybliżyć rozwiązanie $x(t)$ na odcinku $[t_0, T]$, na którym $x \in C^k([t_0, T])$, w dyskretnych punktach czasu:

$$t_n \equiv t_n^h = t_0 + n\,h \qquad h > 0.$$

Często będziemy opuszczali indeks h, o ile to nie będzie powodowało niejasności. Wartość rozwiązania w punkcie t_n^h, czyli $x(t_n)$ będzie przybliżana przez x_n^h, spełniające odpowiedni schemat. Wygodnie jest też oznaczać $f(t_n^h, x_n^h) = f_n^h$. Górny indeks h będziemy często opuszczali, jeśli h będzie ustalone.

Tak więc otwarty schemat Eulera możemy zapisać jako:

$$x_{n+1} = x_n + h\, f_n \qquad n > 0 \qquad x_0 = x(t_0), \tag{3.4}$$

a zamknięty schemat Eulera możemy zapisać jako:

$$x_{n+1} = x_n + h\, f_{n+1} \qquad n > 0 \qquad x_0 = x(t_0). \tag{3.5}$$

Kolejnym wyprowadzeniem otwartego schematu Eulera (3.4) jest obcięcie rozwinięcia szeregu Taylora rozwiązania:

$$x(t+h) = x(t) + \frac{dx}{dt}(t)\,h + \frac{1}{2}\frac{d^2x}{dt^2}(t)\,h^2 + \dots \tag{3.6}$$

Zostawiamy tylko pierwsze dwa człony i otrzymujemy

$$x(t+h) \approx x(t) + \frac{dx}{dt}(t)\,h = x(t) + f(x(t),t)\,h$$

czyli wstawiając x_n za przybliżenie $x(t_n)$, a x_{n+1} za przybliżenie $x(t_n+h) = x(t_{n+1})$ otrzymujemy znów otwarty schemat Eulera (3.4). Analogicznie możemy wyprowadzić zamknięty schemat Eulera (3.5) rozwijając rozwiązanie w t dla $h < 0$.

Jeszcze innym intuicyjnym wyprowadzeniem schematu otwartego Eulera jest *podążanie za polem wektorowym*. Jak wiemy, wykresem rozwiązania równania różniczkowego jest krzywa styczna do zadanego pola wektorowego $f(t,x)$ spełniająca odpowiedni warunek początkowy. Zatem znając rozwiązanie przybliżone dla t_n tzn. mając x_n, możemy wyznaczyć x_{n+1}, przybliżenie rozwiązania $x(t_{n+1})$, biorąc poprawkę w kierunku pola wektorowego tzn.:

$$x_{n+1} = x_n + h\,f(x_n, t_n).$$

Czyli znów otrzymujemy otwarty schemat Eulera.

Zadajmy pytanie, czy takie schematy są wystarczająco dokładne. Czy one działają stabilnie na dłuższych odcinkach czasu, na których istnieje rozwiązanie?

Sprawdźmy, co się dzieje dla modelowego zadania:

$$\frac{dx}{dt} = a\,x \qquad x(0) = 1,$$

którego rozwiązaniem jest $x(t) = e^{at}$.

Otwarty schemat Eulera daje nam ciąg:

$$x_n^h = x_{n-1} + h\,a\,x_{n-1} = (1 + h\,a)x_{n-1} = (1 + h\,a)^n.$$

Ustalmy $t = h\,n$, czyli $h = t/n$. Wtedy

$$x_n^h = (1 + h\,a)^n = (1 + t\,a/n)^n \to e^{at} \qquad n \to \infty.$$

Dla zamkniętego schematu Eulera otrzymujemy analogicznie:

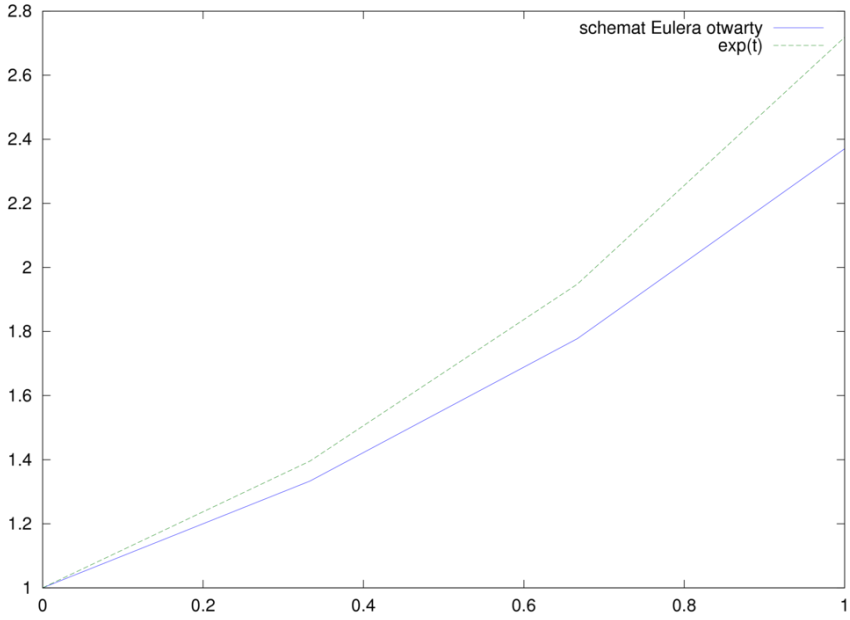
$$x_n^h = x_{n-1}^h + h\,a\,x_n^h,$$

czyli

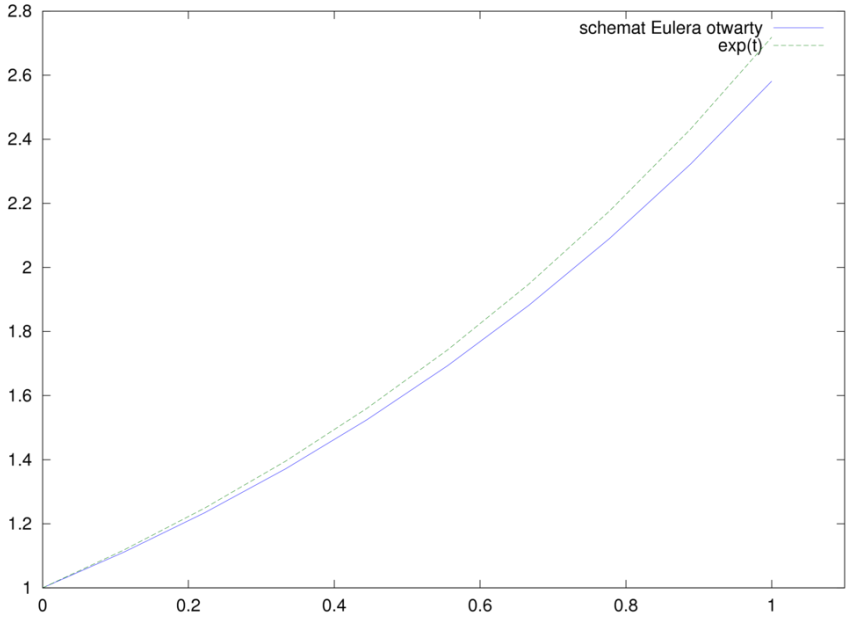
$$x_n = (1 - h\,a)^{-1}\,x_{n-1} = (1 - a\,h)^{-n} = \frac{1}{(1 - a\,t/n)^n} \to \frac{1}{e^{-a\,t}} = e^{a\,t}.$$

Popatrzmy jak te dwa schematy działają (w praktyce) na wykresach dla $a = 1$ i $x(0) = x_0 = 1$, por. rysunki 3.1 - 3.4 dla otwartego schematu Eulera i rysunki 3.5 - 3.8 dla zamkniętego schematu Eulera.

Zauważmy, że wykres rozwiązania ze schematu Eulera otwartego jest poniżej wykresu dokładnego rozwiązania, a dla schematu zamkniętego - powyżej, co widać lepiej na rysunku 3.9.



Rysunek 3.1. Otwarty Euler - cztery punkty.

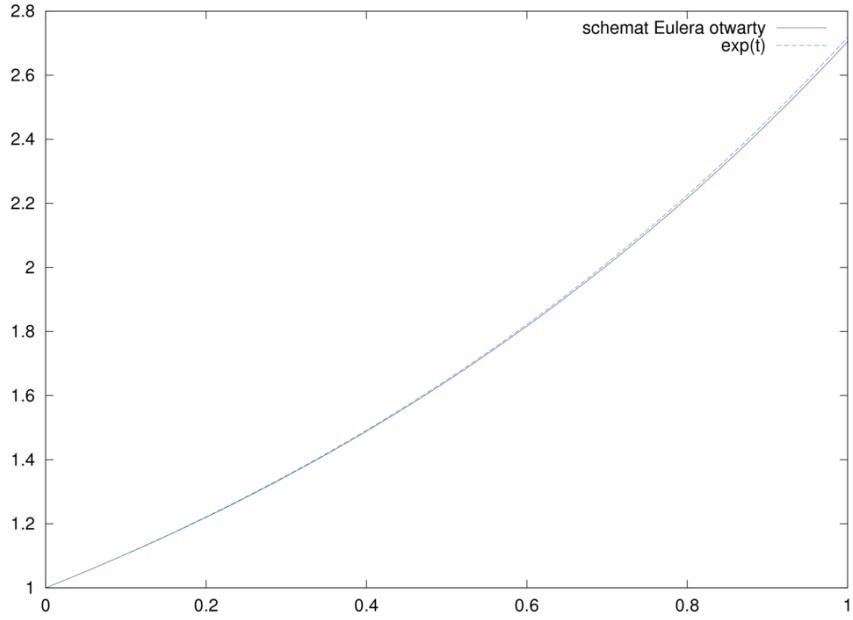


Rysunek 3.2. Otwarty Euler - 10 punktów.

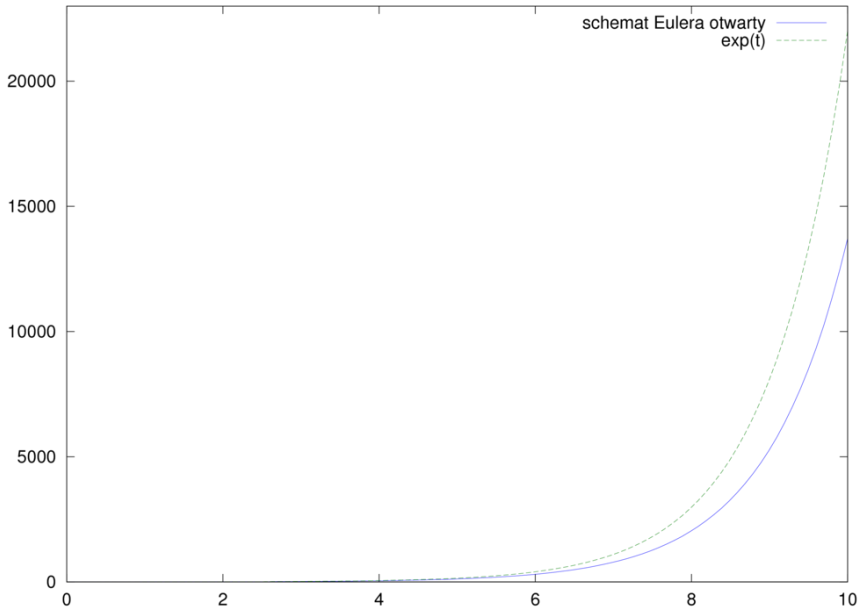
Popatrzmy na przypadek dwuwymiarowy. Weźmy modelowe zadanie wahadła. Dla małych prędkości możemy przyjąć, że $\sin(x) \approx x$, stąd otrzymujemy równanie liniowe ze stałymi współczynnikami (zlinearyzowane równanie wahadła):

$$\frac{d^2x}{dt^2} = -a\,x,$$

gdzie x to prędkość kątowa, a $a = g/l > 0$ dla g wartości przyspieszenia ziemskiego i l długości wahadła.



Rysunek 3.3. Otwarty Euler - 100 punktów.



Rysunek 3.4. Otwarty Euler - na [0,100], 100 punktów.

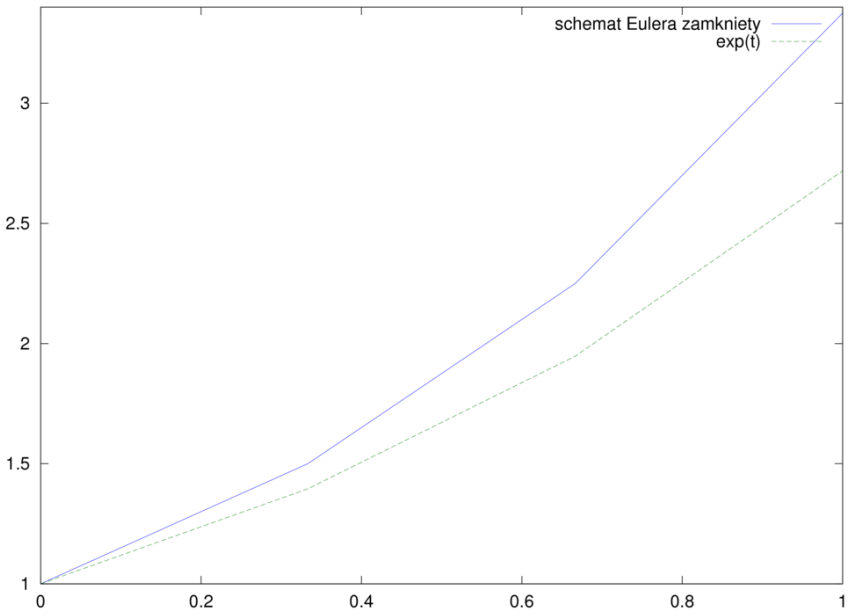
Zapisując to równanie jako układ dwóch równań pierwszego rzędu otrzymujemy:

$$\begin{aligned} \frac{dx}{dt} &= y \\ \frac{dy}{dt} &= -a\,x. \end{aligned}$$

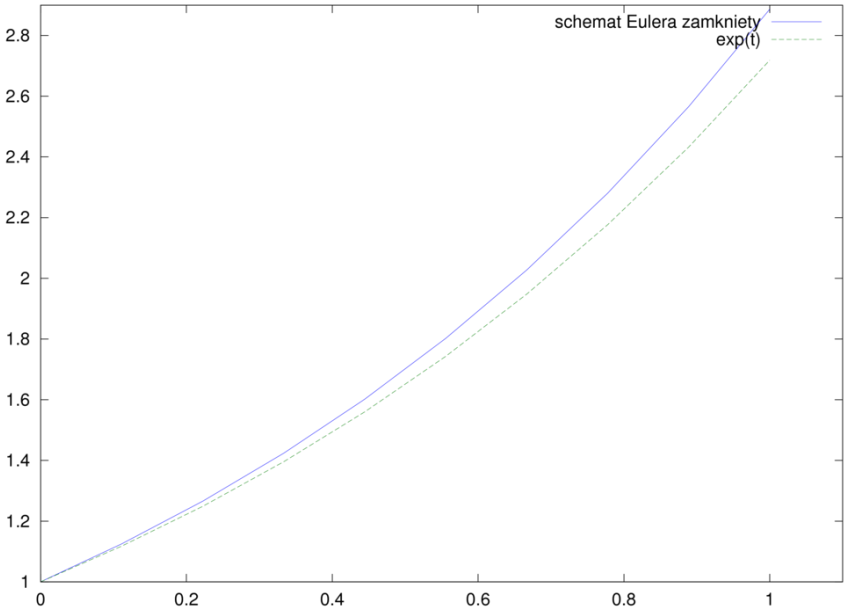
Przyjmijmy, że $a = 1$.

Znamy rozwiązania:

$$x(t) = c_1\,\sin(t) + c_2\,\cos(t),$$



Rysunek 3.5. Zamknięty Euler - cztery punkty.



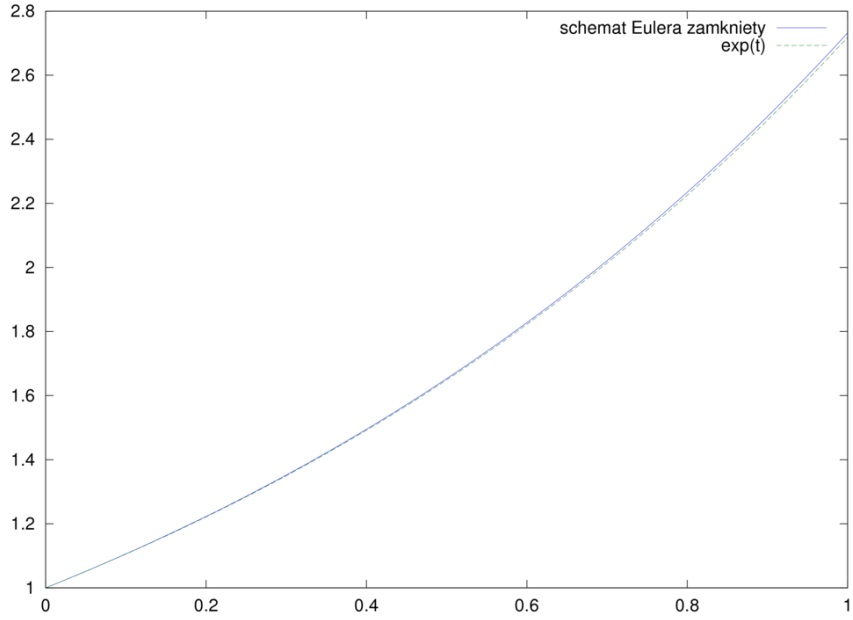
Rysunek 3.6. Zamknięty Euler - 10 punktów.

czyli trajektorie rozwiązania zawarte są w okręgach.

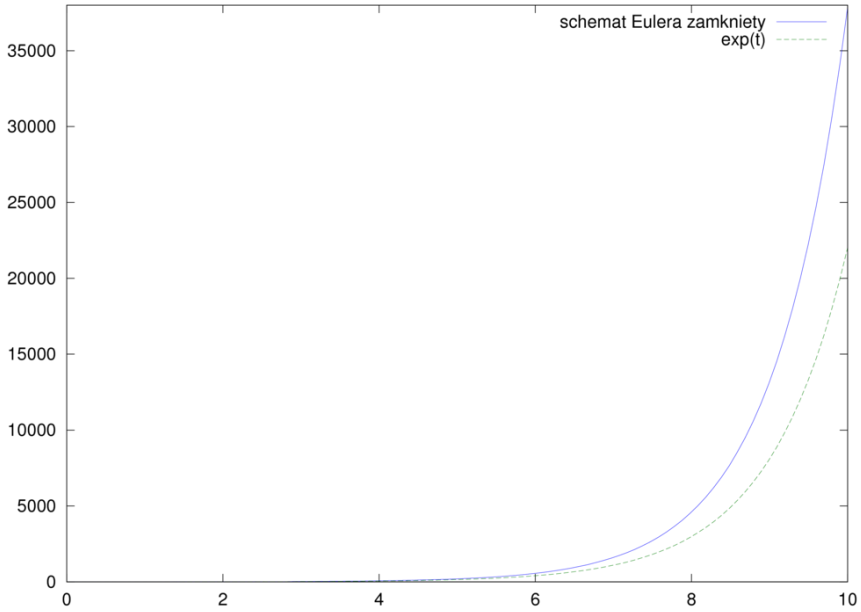
A teraz zastosujmy otwarty schemat Eulera do tego równania z warunkiem początkowym $(x(0), y(0))^T = (0, 1)^T$, którego rozwiązaniem jest $x(t) = \sin(t)$ z $y(t) = \cos(t)$:

$$\begin{cases} x_{n+1} = x_n + h\, y_n \\ y_{n+1} = y_n - h\, x_n \end{cases} \qquad n = 0, 1, \ldots, N$$

dla ustalonego $h > 0$ i $T = N\, h$.



Rysunek 3.7. Zamknięty Euler - 100 punktów.



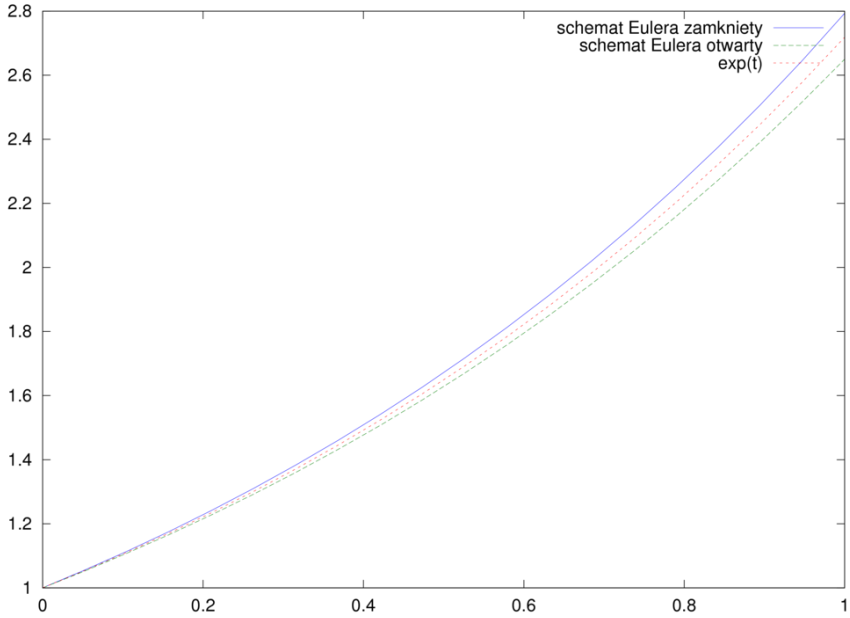
Rysunek 3.8. Zamknięty Euler - na [0,100], 100 punktów.

Zatem: $x_n \approx x(t_n) = \sin(n\,h)$, a $y_n \approx y(t_n) = \cos(n\,h)$ z $x_0 = 0$ i $y_0 = 1$
Dla zamkniętego schematu Eulera jest analogicznie:

$$\begin{cases} x_{n+1} = x_n + h\,y_{n+1} \\ y_{n+1} = y_n - h\,x_{n+1} \end{cases} \qquad n = 0, 1, \ldots, N,$$

czy równoważnie

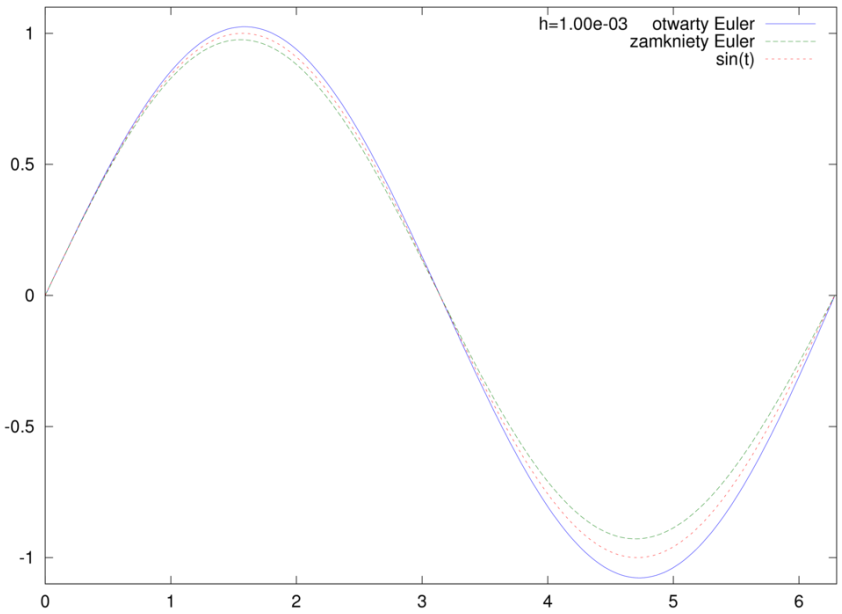
$$\begin{cases} x_{n+1} - h\,y_{n+1} = x_n \\ y_{n+1} + h\,x_{n+1} = y_n \end{cases} \qquad n = 0, 1, \ldots, N$$



Rysunek 3.9. Schematy Euler - na [0,1], 20 punktów.

czyli w każdym kroku dla ustalonego n musimy rozwiązać układ dwóch równań liniowych.

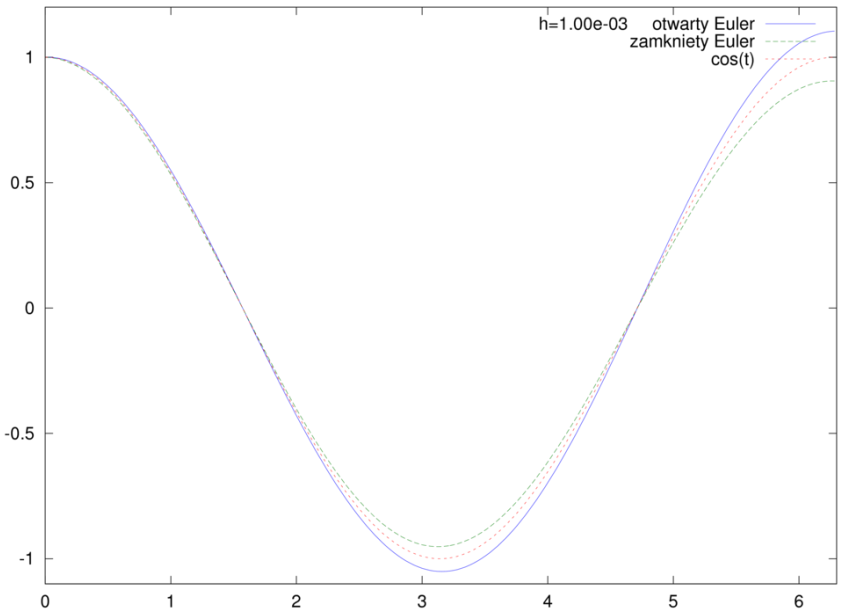
Popatrzmy teraz na rysunki 3.10 - 3.12.



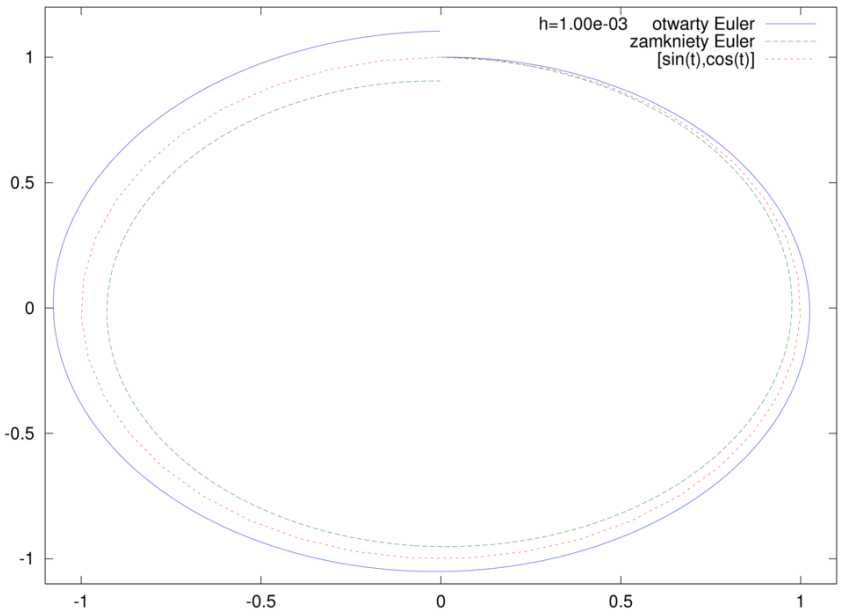
Rysunek 3.10. Schematy Eulera dla równania 2-wymiarowego - rozwiązanie.

Widać, że mimo małego kroku rzędu ($h = 1e-2$) wyniki są wyraźnie gorsze niż w przypadku skalarnym, mimo że wyjściowe równanie różniczkowe jest liniowe.

Rozważmy wyjściowe równanie wahadła, por. Przykład 2.5. Znów przyjmijmy, że $g = l$ i warunek początkowy $x(0) = 0$ i $y(0) = 1$. Wtedy schematy Eulera przybierają odpowiednio formę:



Rysunek 3.11. Schematy Eulera dla równania 2-wymiarowego - pochodna rozwiązania.



Rysunek 3.12. Schematy Eulera dla równania 2-wymiarowego - trajektoria.

schemat otwarty Eulera:

$$\begin{cases} x_{n+1} = x_n + h\, y_n \\ y_{n+1} = y_n + h\, \sin(x_n) \end{cases} \qquad n = 0, 1, \ldots, N$$

z $x_0 = x(0) = 0$ i $y_0 = y(0) = 1$. Znając x_n, y_n otrzymujemy natychmiast wzór na x_{n+1}, y_{n+1}.

W przypadku schematu zamkniętego Eulera:

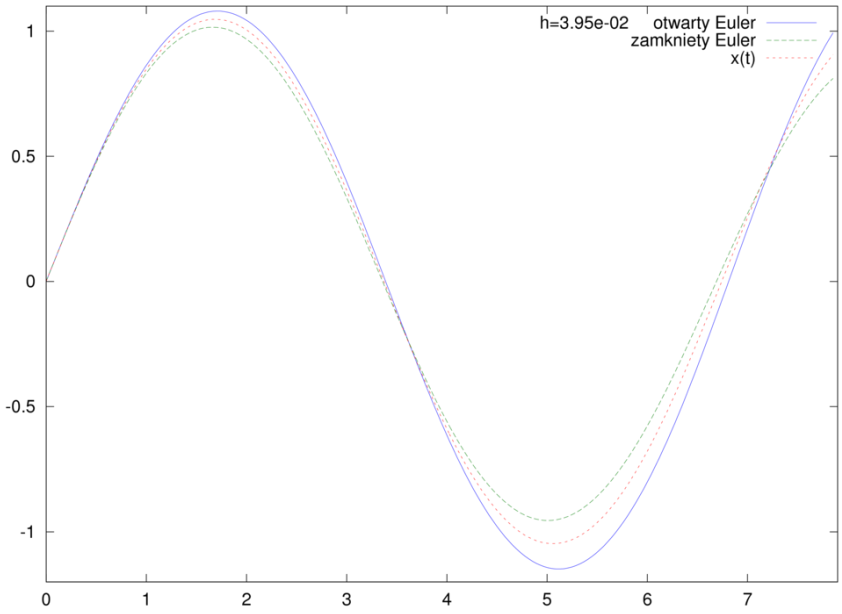
$$\begin{cases} x_{n+1} = x_n + h\, y_{n+1} \\ y_{n+1} = y_n - h\, \sin(x_{n+1}) \end{cases} \qquad n = 0, 1, \ldots, N$$

z $x_0 = x(0) = 0$ i $y_0 = y(0) = 1$, musimy w każdym kroku rozwiązać układ równań nieliniowych:

$$\begin{cases} x_{n+1} - h\, y_{n+1} = x_n \\ y_{n+1} + h\, \sin(x_{n+1}) = y_n \end{cases} \qquad n = 0, 1, \ldots, N$$

Im h jest bliższe zera, tym układ jest łatwiejszy do rozwiązania.

Można pokazać, że rozwiązanie wyjściowego równania ma trajektorie okresowe, co potwierdza wykres na rysunku 3.15 (tu wyliczony przy pomocy dużo dokładniejszego schematu niż schematy Eulera). W kolejnych rysunkach 3.10- 3.13 - prezentujemy przybliżone rozwiązania dla nieliniowego równania wahadła, otrzymane przy pomocy obu schematów Eulera.



Rysunek 3.13. Schematy Eulera dla równania 2-wymiarowego - rozwiązanie.

3.3.1. Absolutna stabilność schematów Eulera

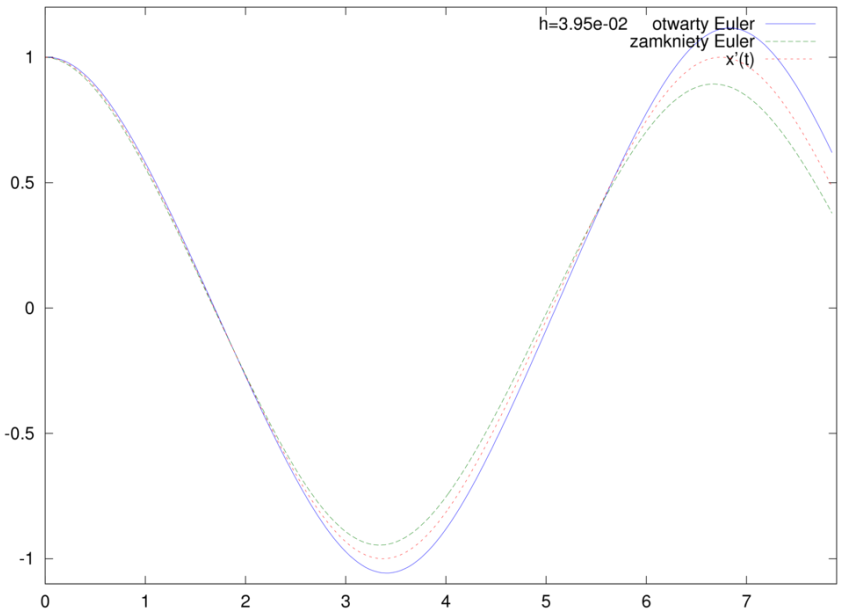
Rozpatrzmy ponownie modelowe zadanie skalarne, ale na długich odcinkach czasu:

$$\frac{dx}{dt} = a\, x, \qquad x(0) = 1 \qquad a < 0.$$

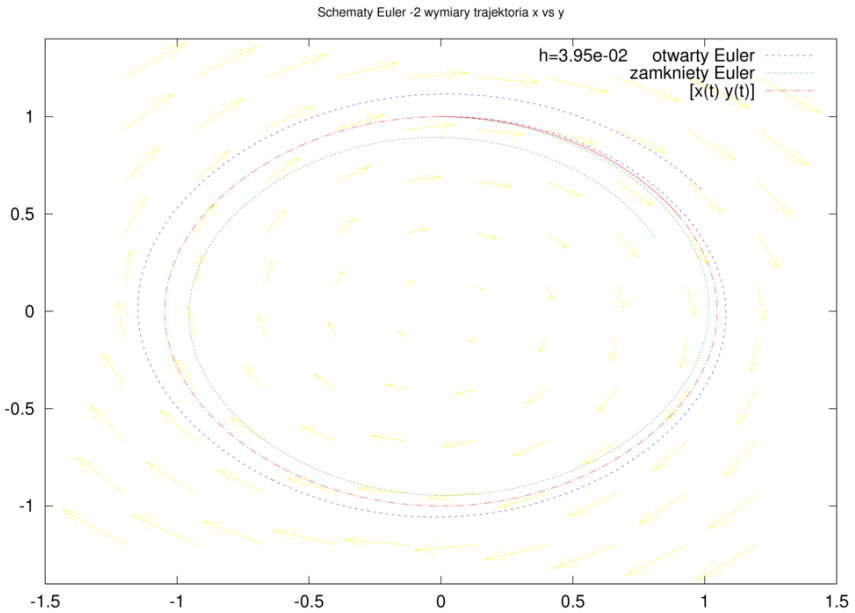
Rozwiązaniem jest $x(t) = \exp(a\, t)$ i wtedy $\lim_{t \to +\infty} x(t) = 0$. Im $|a|$ większe, tym rozwiązanie szybciej zbiega do zera.

Rozpatrzmy teraz zastosowanie otwartego i zamkniętego schematu Eulera do rozwiązania tego zagadnienia. Dla otwartego schematu Eulera wiemy już, że:

$$x_n = (1 + a\, h)^n.$$



Rysunek 3.14. Schematy Eulera dla równania 2-wymiarowego - pochodna rozwiązania.



Rysunek 3.15. Schematy Eulera dla równania 2-wymiarowego - trajektoria.

Zauważmy, że przy ustalonym h ciąg przybliżeń x_n jest dodatni i zbiega do zera dla $n \to +\infty$ o ile zachodzi warunek:

$$h < -1/a.$$

W przypadku gdy parametr a jest ujemny i o dużym module, warunek ten wymusza to, że musimy wziąć bardzo małe h, aby otrzymać schematem otwartym Eulera rozwiązanie przybliżone, które jest dodatnie i malejące do zera, czyli zachowujące się jak rozwiązanie zagadnienia początkowego: $\exp(a\,t)$.

Natomiast dla zamkniętego schematu Eulera widzimy, że:

$$x_n = (1 - a\,h)^{-n}.$$

Otrzymujemy wtedy, że dla dowolnego $a < 0$ zachodzi $x_n > 0$ i $x_n \to 0$ dla $n \to +\infty$, czyli nie otrzymujemy żadnego ograniczenia na krok h, co jest istotne, jeśli chcemy rozwiązywać równanie na długim odcinku czasu.

Schemat zamknięty Eulera można uznać za lepszy od schematu otwartego dla tego zagadnienia dla ujemnego a o bardzo dużym module, szczególnie na długim odcinku czasu, ponieważ nie wymusza żadnych ograniczeń na krok h. Wrócimy do tego problemu w następnych rozdziałach.

3.4. Zadania

Ćwiczenie 3.1. Czy rozwiązanie $y(x)$ zagadnienia początkowego:

$$\frac{dy}{dx} = x^{2/3} \qquad y(0) = 0.$$

jest wyznaczone jednoznacznie? Znajdź wszystkie rozwiązania $y(x)$ tego zagadnienia początkowego. *Wskazówka. Jest to równanie o zmiennych rozdzielonych (autonomiczne)* $\frac{dy}{dx} = f(y)g(x)$ *z warunkiem początkowym* $y(x_0) = y_0$*, więc w postaci uwikłanej rozwiązanie ma postać* $\int_{y_0}^{y} 1/f(y)dy = \int_{x_0}^{x} g(x)dx$.

Ćwiczenie 3.2 (laboratoryjne)**.** Zaimplementuj w octavie otwarty schemat Eulera i zastosuj go do równania:

$$\frac{dy}{dx} = x^{2/3} \qquad y(0) = 0.$$

na różnych odcinkach czasu np. $[0, 1]$ lub $[0, 10]$ i różnych wartości h np. $h = 1e-1, 1e-2, 1e-4$. Zmniejszając h sprawdź, czy ten schemat znajdzie rozwiązanie różne od zera. Następnie weź przybliżenie startowe na poziomie błędu zaokrągleń np. $x_0 = 10^{-16}$ i sprawdź, jakie schemat znajduje rozwiązania; w szczególności, czy są one różne od zera.

Ćwiczenie 3.3. Rozpatrzmy równanie różniczkowe zwyczajne liniowe jednorodne rzędu n o stałych współczynnikach:

$$\frac{d^n x}{dt^n}(t) + a_{n-1}\frac{d^{n-1}x}{dt^{n-1}}(t) + \ldots + a_0 x(t) = 0.$$

1. Poprzez podstawienie $x_k(t) = \frac{d^k x}{dt^k}$ $k = 0, \ldots, n-1$ sprowadź to równanie do równania liniowego jednorodnego ze stałą macierzą:

$$\frac{dx}{dt} = A\,\vec{x},$$

2. Znajdź wielomian charakterystyczny A oraz dla $n = 2$ postać Jordana tej macierzy w zależności od tego, jakie wartości własne ma A,
3. Przy założeniu, że A ma n jednokrotnych wartości własnych rzeczywistych, znajdź $e^{A\,t}$ i dla $n = 2$ znajdź rozwiązanie zadania początkowego dla tego równania z warunkami początkowymi: $\frac{d^k x}{dt^k}(0) = b_k \qquad k = 0, \ldots, n-1.$

Ćwiczenie 3.4 (częściowo laboratoryjne)**.** Dla $n = 2$ i macierzy A kolejno $[a, 1; 0, a]$, $[a, 0; 0, b]$, $[a, -b; b, a]$ dla rożnych wartości parametrów a, b, np. $a = 1, b = 10$, naszkicuj na kartce portrety fazowe (wykresy trajektorii) równania jednorodnego:

$$\frac{dx}{dt} = A\,\vec{x}$$

w otoczeniu zera. Naszkicuj pole wektorowe na ekranie korzystając z funkcji octave'a **quiver**() i portrety fazowe z pomocą funkcji lsode().

Ćwiczenie 3.5 (laboratoryjne)**.** Zaimplementuj w octavie otwarty schemat Eulera i zastosuj go do równania $\frac{dy}{dx} = a\,y$ z $y(0) = 1$ dla różnych wartości parametru a np. $a = -1e-3, -100, -1, 1, 10$. Narysuj na monitorze wykresy przybliżonych rozwiązań razem z wykresem rozwiązania dokładnego $y(t) = \exp(a\,t)$.

Ćwiczenie 3.6 (częściowo laboratoryjne)**.** Rozpatrzmy równanie $\frac{dy}{dx} = \begin{pmatrix} -20 & 1 \\ -21 & 1 \end{pmatrix} y$. Policz wartości własne macierzy $A = \begin{pmatrix} -20 & 1 \\ -21 & 1 \end{pmatrix}$ i porównaj z wynikiem obliczonym w octavie z użyciem odpowiedniej funkcji np. **eig**(). Znajdź rozwiązanie ogólne tego równania. Przy pomocy otwartego schematu Eulera i funkcji octave'a lsode() rozwiąż to równanie z $y(0) = (1, 1)^T$ na odcinku $[0, 100]$ z $h = 0.1$. Porównaj wyniki rysując wykresy na ekranie obu rozwiązań i rozwiązania dokładnego, które należy też wyznaczyć.

Wskazówka. Rozwiązanie ogólne - to $\exp(A\,t)\,c$*, gdzie c wektor stałych a funkcja* ***expm()*** *octave'a pozwala obliczyć eksponent macierzy.*

Ćwiczenie 3.7 (częściowo laboratoryjne)**.** Udowodnij, że przybliżenia rozwiązania układu równań $\frac{dx}{dt} = y; \frac{dy}{dt} = -x$ z $x(0) = 1, y(0) = 0$, otrzymane za pomocą otwartego (lub zamkniętego) schematu Eulera, mają normę drugą zbieżną do jeden, tzn. $\sqrt{(x_n)^2 + (y_n)^2}$ zbiegają do jeden, dla ustalonego czasu $t = n\,h$ z h dążącym do zera. Zaimplementuj oba schematy Eulera dla tego równania w octave (w przypadku zamkniętego schematu Eulera użyj operatora backslash: w każdym kroku czasowym do rozwiązania odpowiedniego układu dwóch równań liniowych). Naszkicuj na ekranie monitora portret fazowy przy pomocy **plot**(), lsode() i obu schematów dla różnych wartości h. Policz wartości normy drugiej rozwiązań otrzymanych przy pomocy tych schematów i lsode() dla ustalonego czasu np. $t = 1$ czy $t = 1000$ i różnych wartości h.

Ćwiczenie 3.8 (laboratoryjne)**.** Naszkicuj na ekranie monitora portrety fazowe równań liniowych $\frac{dy}{dx} = Ay = [a, b; c, d]y$ dla macierzy A o różnych postaciach Jordana przy pomocy **plot**(), lsode().

4. Metody dla równań różniczkowych zwyczajnych - rząd schematów

W tym rozdziale zajmiemy się pewnymi własnościami schematów dla równań różniczkowych zwyczajnych. W szczególności przedstawimy pojęcie rzędu schematu oraz zdefiniujemy, co oznacza zbieżność schematu z odpowiednim rzędem.

4.1. Kilka kolejnych schematów

Można postawić pytanie, czy istnieją schematy o wyższej dokładności niż schematy Eulera. Okazuje się, że tak jest i w tym rozdziale przedstawimy kolejne schematy, które dokładniej przybliżają rozwiązanie wyjściowego problemu różniczkowego.

Dość niska dokładność schematów Eulera, którą zaobserwowaliśmy w eksperymentach z rozdziału 3 wynika z tego, że pochodną rozwiązania przybliżyliśmy najprostszym ilorazem różnicowym. W schematach Eulera przybliżamy pochodną poprzez iloraz różnicowy dla parametru $h > 0$ i otrzymujemy:

$$\left| \frac{x(t+h) - x(t)}{h} - \frac{dx}{dt}(t) \right| = O(h) \tag{4.1}$$

o ile x ma ciągłą drugą pochodną w otoczeniu t.

Jeśli x jest bardziej regularna, to pochodną można przybliżyć dokładniej, np. poprzez iloraz różnicowy centralny (pochodna różnicowa centralna)

$$\left| \frac{x(t+h) - x(t-h)}{2h} - \frac{dx}{dt}(t) \right| = O(h^2). \tag{4.2}$$

Dowód pozostawiamy jako zadanie.

Otrzymujemy w ten sposób:

$$x_{n+1} = x_{n-1} + 2\,h\,f_n \tag{4.3}$$

czyli schemat kroku środkowego (midpoint) dla (3.1) .

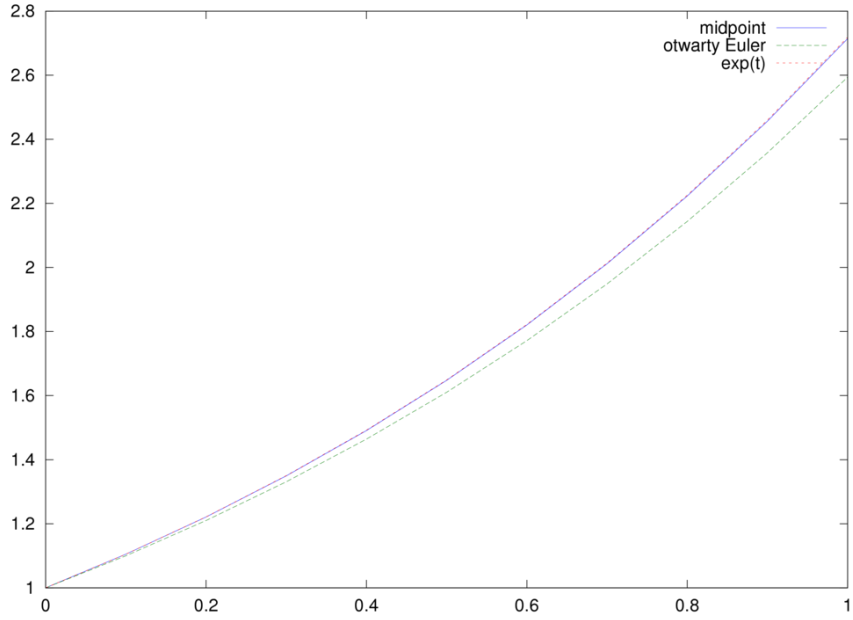
Schemat midpoint, czyli kroku środkowego, jest dwu-krokowy, tzn. że aby obliczyć x_{n+1} musimy znać x_n i x_{n-1}, czyli trzeba znać x_0 i x_1.

Policzmy przy pomocy tego schematu rozwiązanie zagadnienia początkowego:

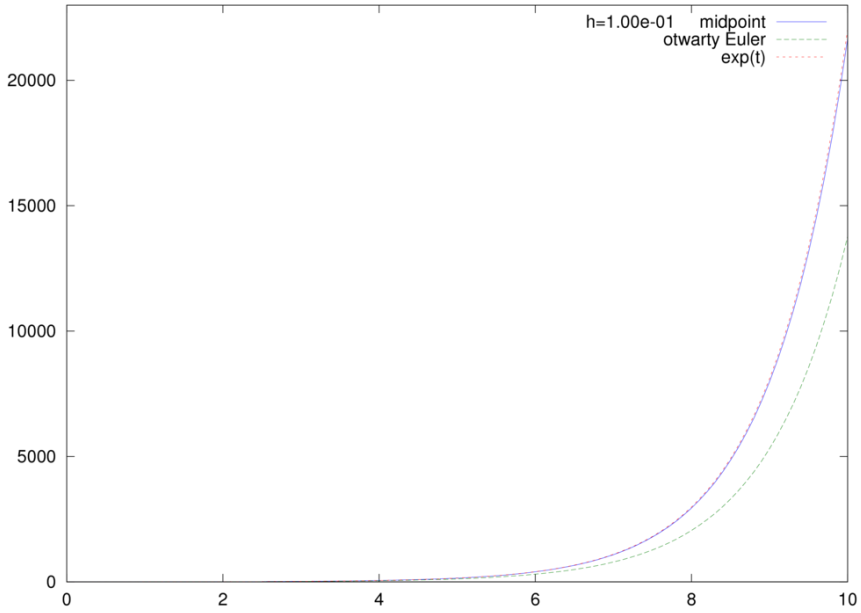
$$\frac{dx}{dt} = a\,x \qquad x(0) = 1$$

Na początek weźmy $a = 1$ i porównajmy z rozwiązaniem; za x_1 do naszych testów schematu midpoint weźmiemy dokładną wartość rozwiązania: $\exp(h)$, por. rysunek 4.1. Wyraźnie dokładniejszym okazuje się schemat midpoint.

Można się zastanowić, co się stanie na dłuższym odcinku czasu, por. rysunek 4.2. Okazuje się, że schemat midpoint dokładniej działa także w tym przypadku.



Rysunek 4.1. Schematy midpoint i Eulera otwarty na odcinku [0,1].



Rysunek 4.2. Schematy midpoint i Eulera otwarty na odcinku [0,10].

Schemat ten nie jest jednak w ogóle używany. W kolejnym rozdziale wyjaśnimy dlaczego.

Inną drogą wprowadzenia nowych schematów jest skorzystanie z rozwinięcia rozwiązania w szereg Taylora: (3.6), tak jak dla schematu Eulera, ale z większą ilością członów. Otrzymujemy

w ten sposób np. schemat Taylora:

$$\begin{aligned} x(t+h) &\approx x(t) + \frac{dx}{dt}(t)\,h + \frac{1}{2}\frac{d^2x}{dt^2}(t)\,h^2 \\ &= x(t) + f(t,x(t))\,h + \frac{h^2}{2}\,(\frac{\partial f}{\partial x}(t,x(t))f(t,x(t)) + \frac{\partial f}{\partial t}(t,x(t))). \end{aligned}$$

Skorzystaliśmy tu z tego, że $\frac{d^2x}{dt^2} = \frac{d}{dt}f(t,x(t)) = \frac{\partial f}{\partial x}(t,x(t))\frac{dx}{dt}(t) + \frac{\partial f}{\partial t}(t,x(t))$.
Schemat Taylora, a dokładniej schemat Taylora rzędu dwa, wygląda następująco:

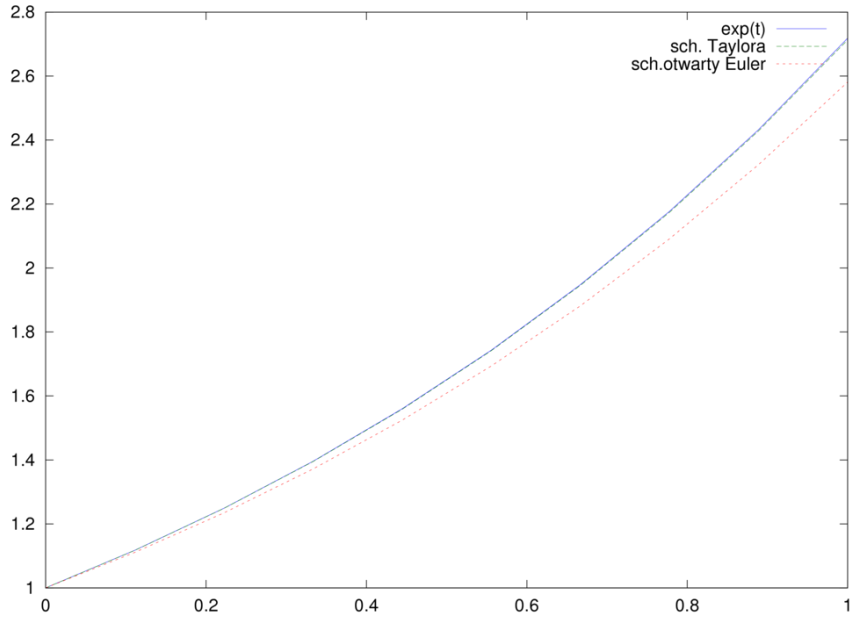
$$x_{n+1} = x_n + h\,f_n + \frac{h^2}{2}\,(\partial_x f_n\,f_n + \partial_t f_n), \tag{4.4}$$

gdzie $\partial_x f_n = \frac{\partial f}{\partial x}(t_n, x_n)$ i $\partial_t f_n = \frac{\partial f}{\partial t}(t_n, x_n)$. W przypadku równania autonomicznego ($f(t,x) = f(x)$) schemat się upraszcza i otrzymujemy:

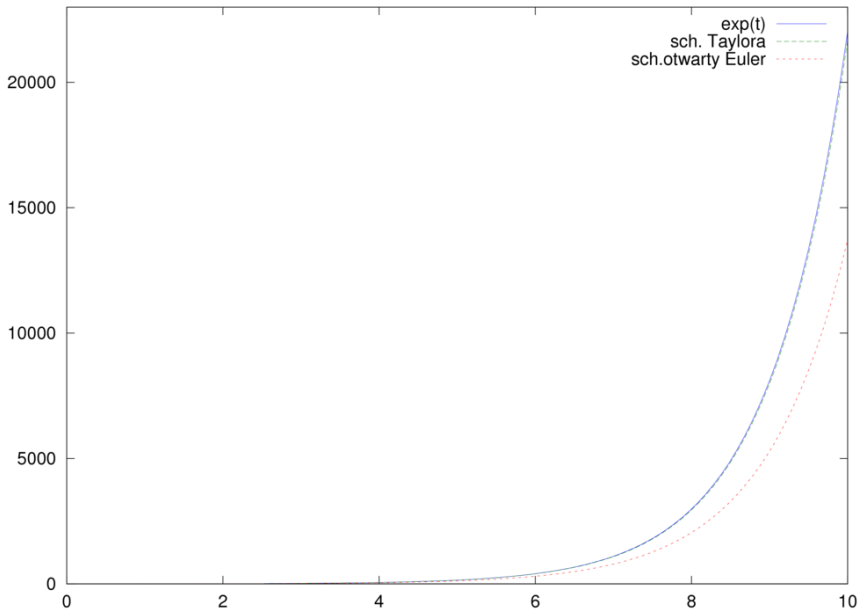
$$x_{n+1} = x_n + h\,f_n + \frac{h^2}{2}\,(\partial_x f_n\,f_n).$$

Proszę zauważyć, że ogólnie $\partial_x f(t,x)$ jest macierzą $m \times m$, a $\partial_t f(t,x)$ jest wektorem wymiaru m, czyli koszt schematu Taylora w przypadku wielowymiarowym dla $m > 1$ jest dość duży. Musimy obliczyć w każdym kroku dwa wektory tzn. f_n i $\partial_t f_n$ oraz macierz $\partial_x f_n$, wymnożyć tę macierz przez f_n i przemnożyć odpowiednie wektory przez h, h^2 i dodać je do siebie. Możemy w ten sposób tworzyć kolejne schematy Taylora o coraz większej dokładności - jeśli f jest funkcją dostatecznie gładką. Będą to schematy coraz droższe, szczególnie w przypadku dużego wymiaru m.

Na rysunkach 4.3 i 4.4 widać, że podobnie jak dla schematu midpoint, schemat Taylora jest dokładniejszy niż schemat Eulera otwarty.



Rysunek 4.3. Rozwiązanie dokładne $y' = y$ z $y(0) = 1$, rozwiązanie schematem Taylora i otwartym schematem Eulera na [0,1] z $h = 0.1$.



Rysunek 4.4. Rozwiązanie dokładne $y' = y$ z $y(0) = 1$, rozwiązanie schematem Taylora i otwartym schematem Eulera na [0,10] z $h = 0.1$.

4.1.1. Zbieżność metod - idea

Błąd schematu (ang. *global error*) np. Eulera otwartego, czy zamkniętego, czy schematu midpoint zastosowanych do rozwiązywania przybliżonego (3.1) możemy zdefiniować dla ustalonego $t \in [t_0, T]$ jako

$$E_h(t) = |x_n^h - x(t)|$$

dla $h = (t - t_0)/n$, x rozwiązania (3.1) na $[t_0, T]$. Zbadajmy, jak zachowuje się błąd $E_h(t)$ wraz ze zmniejszaniem h w ustalonym t. W szczególności, czy maleje do zera.

Popatrzmy, co pokazują eksperymenty - zastosowaliśmy otwarty schemat Eulera z różnymi krokami do policzenia przybliżenia rozwiązania równania $dx/dt = x$ z $x(0) = 1$ dla czasu $t = 1$, czyli znamy dokładną wartość rozwiązania $x(1) = e$. Ustaliliśmy $h_0 = 0.5$, a następnie kolejno je połowiliśmy tzn. $2^{-1}h_0, \ldots, 2^{-5}h_0$. Wyniki są w tabeli 4.1.

h	*otwarty Euler*	*midpoint*	*Taylor*
$5.0e-01$	$-4.7e-01$	$-7.8e-02$	$-7.0e-02$
$2.5e-01$	$-2.8e-01$	$-2.3e-02$	$-2.4e-02$
$1.2e-01$	$-1.5e-01$	$-6.4e-03$	$-6.6e-03$
$6.2e-02$	$-8.0e-02$	$-1.7e-03$	$-1.7e-03$
$3.1e-02$	$-4.1e-02$	$-4.3e-04$	$-4.4e-04$
$1.6e-02$	$-2.1e-02$	$-1.1e-04$	$-1.1e-04$
$7.8e-03$	$-1.1e-02$	$-2.7e-05$	$-2.8e-05$

Tabela 4.1. Błąd dla schematów: otwartego, Eulera, schematu midpoint i schematu Taylora, przybliżających rozwiązanie $dx/dt = x$ z $x(0) = 1$ dla $t = 1$ czyli $\exp(1)$.

Widać, że dla schematu Eulera błąd dla zmniejszonego dwukrotnie h maleje dwukrotnie co

sugeruje, że błąd zachowuje się jak $O(h)$, gdy dla schematów midpoint i schematu Taylora błąd maleje czterokrotnie, czyli zachowuje się jak $O(h^2)$.

W schemacie midpoint przybliżamy pochodną różnicą centralną, dla której zachodzi:

$$(x(t+h) - x(t-h))/(2\,h) = \frac{dx}{dt}(t) + O(h^2)$$

dla dostatecznie gładkiej funkcji, a w przypadku otwartego schematu Eulera - zwykłym ilorazem różnicowym

$$(x(t+h) - x(t))/(h) = \frac{dx}{dt}(t) + O(h).$$

Przy konstrukcji schematu Taylora wykorzystujemy więcej członów z rozwinięcia rozwiązania w szereg Taylora (3.6). Każdy dodatkowy człon z szeregu Taylora powinien podwyższyć dokładność danego schematu.

Dlatego też wprowadza się pojęcie rzędu lokalnego błędu schematu (ang. *local truncation error*), czyli rzędu schematu. Badamy lokalny błąd schematu względem parametru h, jeśli wstawimy za x_n dokładną wartość rozwiązania $x(t_n)$. Najpierw zdefiniujmy samo pojęcie schematu rozwiązywania (3.1), potem zbieżności schematu i rzędu schematu.

Definicja 4.1. Schematem k krokowym rozwiązywania zadania początkowego (3.1) ze stałym krokiem $h > 0$ na odcinku $[t_0, T]$) nazywamy równanie różnicowe:

$$x_n = \Phi(h, t_n, x_{n-k} \ldots, x_{n-1}, x_n) \qquad n \geqslant k \tag{4.5}$$

z warunkami startowymi $x_0, \ldots, x_{k-1}$ dla $t_n = t_0 + n\,h$. Jeśli Φ nie zależy od x_n, to schemat nazywamy otwartym (ang. *explicit*). W przeciwnym razie - schemat nazywamy zamkniętym (ang. *implicit*).

Schematy konstruujemy tak, aby dla ustalonego h zachodziło $x_n \approx x(t_n)$.

Definicja 4.2. Niech $x \in C^1([t_0, T))$ rozwiązania zagadnienia początkowego (3.1). Błąd schematu k krokowego postaci (4.5) dla $t = t_0 + n\,h \in [t_0, T]$ definiujemy jako

$$E_h(t) = |x_n^h - x(t)| \qquad (t = t_0 + n\,h),$$

a błąd globalny (ang. *global error*) na $[t_0, T]$ jako

$$E_h = \max_{n=0,\ldots,N} E_h(t_n^h)$$

dla $N = (T - t_0)/h$. Schemat jest zbieżny na $[t_0, T]$, jeśli

$$E_h \to 0 \qquad h \to 0,$$

a jest zbieżny z rzędem p (ang. *convergent with order p*) (rząd błędu globalnego wynosi p), jeśli dodatkowo

$$E_h \leqslant C\,h^p$$

dla pewnej stałej $C > 0$ niezależnej od h (zazwyczaj zależnej od rozwiązania x (3.1) i $T - t_0$).

Definicja 4.3. Niech $x \in C^1([t_0, T))$ będzie rozwiązaniem zagadnienia początkowego (3.1). Dla parametru $h > 0$ i schematu k krokowego postaci (4.5) błąd lokalny (ang. *local truncation error*) definiujemy jako

$$e_h = \max_{t \in [t_0, T - k\,h]} |x(t + k\,h) - \Phi(t, \ldots, t + k\,h, x(t), \ldots, x(t + k\,h))|.$$

Definicja 4.4. Schemat (4.5) jest rzędu p (ang. *local truncation error is of order p*), jeśli dla $x \in C^{p+1}([t_0, T))$ rozwiązania zagadnienia początkowego (3.1) zachodzi

$$e_h \leqslant C\, h^{p+1}$$

dla pewnej dodatniej stałej C niezależnej od h.

Dla otwartego schematu Eulera lokalny błąd schematu jest równy:

$$e_h = \max_{t\in[t_0,T-h]} |x(t+h) - x(t) - hf(t, x(t))|.$$

Z rozwinięcia w szereg Taylora widzimy, że:

$$e_h = h \max_{t\in[t_0,T-h]} \left| \frac{x(t+h) - x(t)}{h} - \frac{dx}{dt}(t) \right| = O(h^2)$$

o ile x rozwiązanie (3.1) jest klasy C^2, czyli schemat ma rząd jeden. Analogicznie można pokazać, że rząd zamkniętego schematu Eulera jest też jeden, a rząd schematów midpoint i Taylora wynosi dwa. Wykazanie tego, pozostawimy jako zadanie.

4.1.2. Schematy Adamsa

Możemy też wyprowadzić schematy korzystając z równoważnej całkowej wersji zagadnienia początkowego (3.1):

$$x(t+h) = x(t) + \int_t^{t+h} dx/dt(s)ds = x(t) + \int_t^{t+h} f(s, x(s))ds. \tag{4.6}$$

To prowadzi do konstrukcji całej rodziny schematów (tzw. schematów Adamsa). Jeśli wprowadzimy siatkę równomierną z krokiem $h > 0$, tzn. wprowadzamy $\{t_k\}_{k=1}^N$ dla $t_k^h \equiv t_k = t_0 + k\, h$, to możemy przybliżyć wartość rozwiązania $x(t_k)$ zastępując w (4.6) całką z jakiejś aproksymacji funkcji f, którą daje się wyliczyć znając wartości $f_k = f(t_k, x_k)$ dla ustalonej ilości $k = n+1, n, n-1, \ldots$, np. $k = n, n-1, n-2$. Wtedy

$$x_{n+1} = x_n + \int_{t_n}^{t_{n+1}} P(s)ds,$$

gdzie $P(s)$ jest jakimś wielomianem przybliżającym $f(s, x(s))$ zdefiniowanym poprzez wartości odpowiednie f_k dla $k \leqslant n+1$.

W przypadku schematów Adamsa, $P(t)$ definiujemy jako odpowiedni wielomian interpolacyjny Lagrange'a dla funkcji f z węzłami w punktach t_{n+j} dla $j = 1, 0, -1, ..$ dla schematu zamkniętego (lub $j = 0, -1, ..$ dla schematu otwartego), spełniający odpowiednie warunki interpolacyjne:

$$P(t_{n+j}) = f_{n+j} = f(t_{n+j}, x_{n+j})$$

dla $p+1$ kolejnych indeksów $j \leqslant 1$ dla schematów Adamsa zamkniętych i $j < 1$ dla schematów Adamsa otwartych. Wtedy otrzymujemy klasę zamkniętych schematów Adamsa-Moultona postaci:

$$x_{n+1} = x_n + \sum_{j=-p+1}^{1} \hat{\beta}_j f_{n+j}$$

lub otwartych schematów Adamsa-Bashfortha:

$$x_{n+1} = x_n + \sum_{j=-p}^{0} \hat{\beta}_j f_{n+j}$$

Przenumerowując indeksy uzyskujemy schemat zamknięty Adamsa-Moultona p krokowy:

$$x_{n+p} = x_{n+p-1} + h \sum_{j=0}^{p} \beta_j f_{n+j}$$

lub otwarty $p+1$ krokowy Adamsa-Bashfortha:

$$x_{n+p+1} = x_{n+p} + h \sum_{j=0}^{p} \beta_j f_{n+j}.$$

Oczywiście w obu przypadkach β_j nie zależą od rozwiązania x, ani od f.

W szczególności dla $p = 1$, $P(t)$ jest wielomianem interpolacyjnym stałym, zdefiniowanym przez wartość w jednym punkcie odpowiednio t_n czy t_{n-1}. Dla

$$p = 1 \qquad P(s) = f_n,$$

otrzymujemy schemat otwarty Eulera:

$$x_{n+1} = x_n + \int_{t_n}^{t_n+h} f_n ds = x_n + h\, f_n$$

Biorąc wartość w punkcie t_{n+1} uzyskujemy schemat zamknięty Eulera. A dla $p = 2$, $P(s)$ jest wielomianem liniowym interpolującym f w punktach t_n i t_{n+1}. Wtedy otrzymujemy schemat trapezów (ang. *trapezoidal scheme*):

$$P(s) = h^{-1}\left((t_{n+1} - s)\, f_n + (s - t_n)\, f_{n+1}\right),$$

czyli

$$x_{n+1} = x_n + \int_{t_n}^{t_n+h} P(s) ds = x_n + 0.5\, h\, (f_n + f_{n+1}). \tag{4.7}$$

Można pokazać, że schemat trapezów jest rzędu dwa.

W przypadku, gdy punkt t_{n+1} czyli f_{n+1} nie jest uwzględniony w definicji $P(s)$ tzn. $\beta_{p+1} = 0$ rozpatrujemy otwarte schematy Adamsa, które też nazywamy schematami Adamsa-Bashforda, np. schemat otwarty Eulera. W przeciwnym przypadku otrzymujemy schematy zamknięte, które nazywamy schematami Adamsa-Moultona: np. schemat zamknięty Eulera lub schemat trapezów.

4.2. Schematy liniowe wielokrokowe

Definicja 4.5. Dla zadania początkowego (3.1) schematem liniowym wielokrokowym (ang. *linear multistep*) - dokładniej k krokowym dla stałego kroku dla $h = \frac{T-t_0}{n}$ nazywamy równanie różnicowe:

$$\sum_{j=0}^{k} \alpha_j x^h_{n+j} = h \sum_{j=0}^{k} \beta_j f^h_{j+n} \qquad n \geqslant 0 \tag{4.8}$$

z $\alpha_k \neq 0$ i $f^h_j = f(t_j, x^h_j)$ dla $t_j = t_0 + j\, h$.

Jeśli $\beta_k \neq 0$, to schemat nazywamy zamkniętym, a w przeciwnym wypadku mówimy o schemacie otwartym.

Jeśli znamy $x_0^h, x_1^h, \ldots, x_{k-1}^h$ to możemy wyliczyć rozwiązanie schematu $x_j^h \approx x(t_j)$ dla $t_j = t_0 + j\,h$ i $j \geqslant k$ (o ile ono istnieje, co w przypadku schematów zamkniętych nie jest oczywiste).

Zgodnie z Definicją 4.4 schemat liniowy k-krokowy ma rząd $p \geqslant 1$ jeśli dla $x \in C^{p+1}([t_0, T])$ rozwiązania zagadnienia (3.1) dla $t \in [t_0, T])$ takich, że $t + k\,h \leqslant T$ lokalny błąd schematu spełnia

$$e_h(t) := \left| \sum_{j=0}^{k} \alpha_j x(t + j\,h) - h \sum_{j=0}^{k} \frac{dx}{dt}(t + j\,h) \right| \leqslant C\,h^{p+1} \tag{4.9}$$

ze stałą niezależną od C, czyli $e_h = O(h^{p+1})$.

Jeśli za x_n weźmiemy wartości rozwiązania w punktach czasu t_n, to błąd schematu wynosi $O(h^{p+1})$ (dla gładkiego rozwiązania).

Oczywiście schematy Adamsa opisane w rozdziale 4.1.2 są szczególnym przypadkiem schematów liniowych wielokrokowych. Tak więc schematy: otwarty i zamknięty Eulera, schemat midpoint, lub schemat trapezów są schematami wielokrokowymi liniowymi - w myśl naszej definicji.

4.3. Schematy jednokrokowe

W tym podrozdziale wprowadzimy pojęcie schematu jednokrokowego:

Definicja 4.6. Dla zadania początkowego (3.1) schematem jednokrokowym dla stałego kroku $h = \frac{T-t_0}{N}$ nazywamy równanie różnicowe:

$$x_{n+1} = x_n + h\,\phi(h, t_n, x_n, x_{n+1}) \qquad n = 0, \ldots, N \tag{4.10}$$

gdzie $t_j = t_0 + j\,h$ a ϕ jest funkcją ciągłą określoną na $[0, H) \times [t_0, T) \times U_{x_0} \times U_{x_0}$ dla U_{x_0} otoczenia x_0. Dodatkowo, jeśli ϕ nie zależy od x_{n+1}, to schemat jednokrokowy nazywamy otwartym, a w przeciwnym wypadku mówimy o schemacie zamkniętym.

W przypadku schematów otwartych możemy wyliczyć x_{n+1} znając x_n, natomiast w przypadku schematów zamkniętych musimy rozwiązać liniowy, bądź nieliniowy układ równań. Do tej pory poznaliśmy dwa schematy jednokrokowe (które zarazem są schematami liniowymi wielokrokowymi) - czyli oba schematy Eulera i schemat trapezów.

Analogicznie do przypadku schematów liniowych wielokrokowych, zgodnie z Definicją 4.4, schemat jednokrokowy ma rząd $p \geqslant 1$, jeśli dla $x \in C^{p+1}([t_0, T))$ rozwiązania zagadnienia (3.1) dla $1 \leqslant p$, $h = \frac{T-t_0}{N}$ i $t \in [t_0, T])$ lokalny błąd schematu spełnia:

$$e_h(t) := |x(t + h) - x(t) - h\,\phi(h, t, \frac{dx}{dt}(t), \frac{dx}{dt}(t + h))| \leqslant C\,h^{p+1},$$

ze stałą C niezależną od t, czyli $e_h = O(h^{p+1})$ dla dostatecznie gładkiego rozwiązania.

4.3.1. Schematy Rungego-Kutty

Podstawową klasą schematów jednokrokowych są tzw. schematy Rungego-Kutty lub - mówiąc krótko - schematy Rungego. Idea ich jest prosta.

Załóżmy, że znamy x_n, i chcemy wyliczyć wartość x_{n+1} ze wzoru uwzględniającego wartość pola wektorowego nie tylko w x_n, ale również w dodatkowym punkcie $\tilde{x}$. Wtedy

$$x_{n+1} = F(h, t_n, x_n, \tilde{x}).$$

Biorąc schemat otwarty Eulera z krokiem $\tilde{h}$ otrzymujemy punkt

$$\tilde{x} = x_n + \tilde{h}\,f(t, x(t)),$$

który, jak wiemy, przybliża $x(t+\tilde{h})$, ale niedokładnie. Możemy policzyć wartość pola wektorowego f w tym punkcie i następnie, wykorzystując wartość $y_n = f(t_n, x_n)$ i $\tilde{y} = f(t_n + \tilde{h}, \tilde{x})$, znaleźć lepsze przybliżenie $x(t_{n+1})$ - czyli np. za przybliżenie pola wektorowego wziąć ważoną średnią obu wartości $b\,y_n + c\tilde{y}$ dla pewnych ustalonych wag b, c. Możliwości jest wiele. Pojawia się pytanie: jak oceniać różne konstrukcje F? Można tak dobierać F, aby rząd schematu był możliwie duży.

Załóżmy, że $\tilde{h} = a\,h$. Wtedy szukamy schematu postaci:

$$x_{n+1} = x_n + b\,h\,f_n + c\,h\,f(t_n + a\,h, x_n + a\,h\,f(t_n, x_n)) \tag{4.11}$$

tak, aby schemat miał maksymalny rząd.

Rozwijamy rozwiązanie x w szereg Taylora:

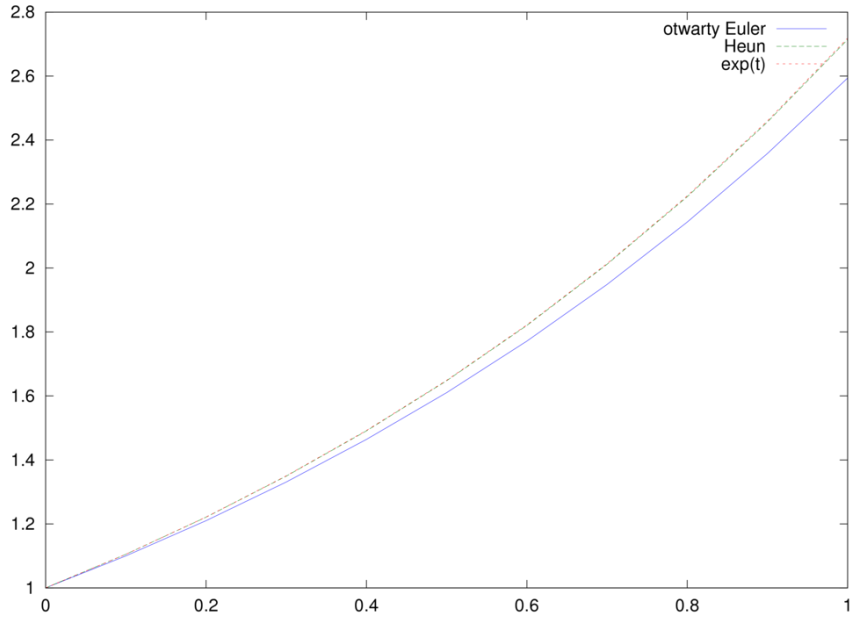
$$x(t+h) - x(t) = h\,\frac{dx}{dt}(t) + 0.5\,h^2\,\frac{d^2x}{dt^2}(t) + O(h^3)$$

i rozwijając ostatni z członów (4.11) w punkcie (t, x) otrzymujemy:

$$\begin{aligned} f(t + a\,h, x + a\,h\,f(t,x)) &= f + a\,h\,f_x\,f + a\,h\,f_t \\ &= f + a\,h(f_x\,f + f_t) \\ &= \frac{dx}{dt} + a\,h\,\frac{d^2x}{dt^2}. \end{aligned}$$

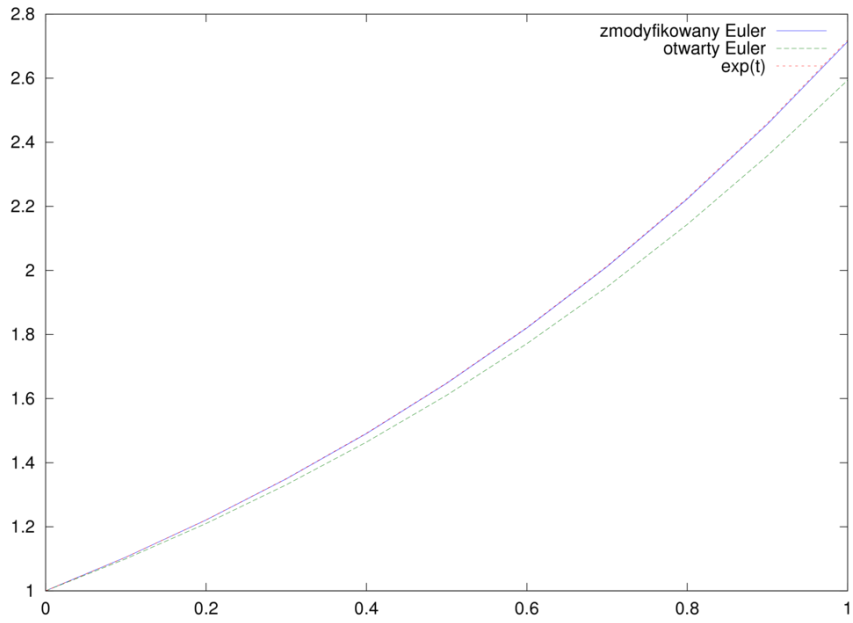
Skorzystaliśmy z tego, że $\frac{d^2x}{dt^2} = \frac{d}{dt}f(t, x(t)) = f_x\,f + f_t$. Zatem, wstawiając dwa ostatnie równania do (4.11) otrzymujemy warunki na to, aby schemat był rzędu dwa:

$$\begin{aligned} b + c &= 1 \\ c\,a &= 0.5 \end{aligned}$$



Rysunek 4.5. Schemat Heuna w porównaniu ze schematem otwartym Eulera na [0,1] z $h = 0.1$.

Tak więc otrzymaliśmy całą rodzinę schematów Rungego-Kutty rzędu dwa, np.:



Rysunek 4.6. Zmodyfikowany schemat Euler w porównaniu ze schematem otwartym Eulera na [0,1] z $h = 0.1$.

1. *Zmodyfikowany schemat Eulera*

$$x_{n+1} = x_n + h\, f\left(t_n + \frac{h}{2}, x_n + \frac{h}{2}\, f_n\right) \tag{4.12}$$

dla $c = 1, b = 0, a = 0.5$,

2. *Schemat Heuna*

$$x_{n+1} = x_n + \frac{h}{2}\left(f_n + f(t_{n+1}, x_n + h\, f_n)\right), \tag{4.13}$$

dla $b = c = 0.5; a = 1$.

Warto zauważyć, że w niektórych publikacjach wszystkie schematy otwarte Rungego-Kutty rzędu dwa nazywane są zmodyfikowanym schematem Eulera.

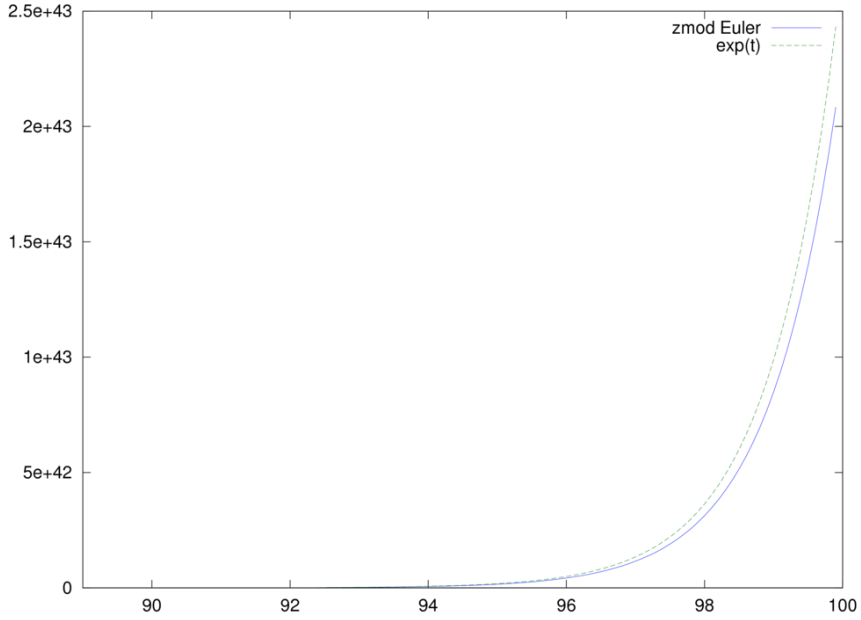
Na rysunkach 4.6 i 4.5 pokazano rozwiązania uzyskane tymi dwoma schematami dla zadania $dx/dt = x$ z $x(0) = 1$ na $[0, 1]$. Widać, że wykresy się pokrywają z wykresem rozwiązania. W porównaniu do otwartego schematu Eulera widzimy znaczącą poprawę. Zobaczmy, co się dzieje na dłuższym odcinku czasu w przypadku schematu Heuna, por. rysunek 4.7.

Graficzne wytłumaczenie zmodyfikowanego schematu Eulera

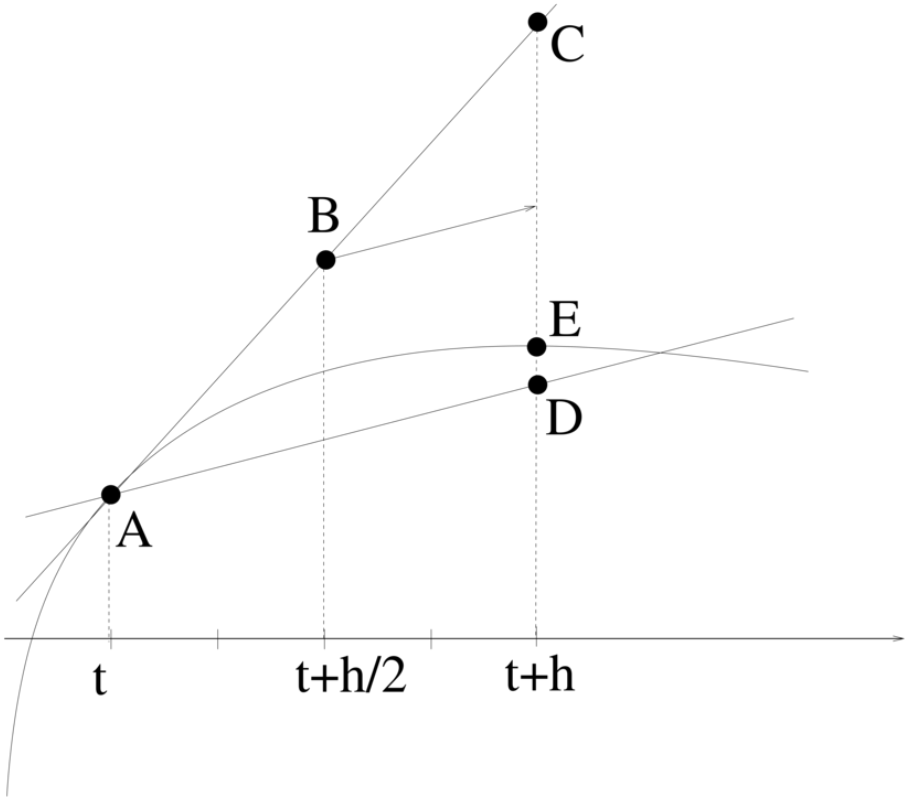
Na rysunku 4.8 zawarliśmy graficzne wytłumaczenie jednego kroku zmodyfikowanego schematu Eulera.

Punkt A - oznacza (t, x), czyli wcześniej obliczone przybliżenie dla czasu t. Chcemy wyznaczyć przybliżenie dla czasu $t+h$. Otwarty schemat Eulera w jednym kroku przyjmuje za różnicę między kolejnymi punktami $h\, f(t, x)$, czyli *idzie w kierunku pola wektorowego* w punkcie t, czyli na naszym rysunku daje to punkt C. Z kolei w zmodyfikowanym schemacie Eulera przyjmujemy za różnicę h pomnożone przez kierunek pola wyznaczonego w dodatkowym pomocniczym punkcie $\hat{x} = x + 0.5\, f_n$ oznaczonym jako B, tzn. *idziemy w kierunku* $f(t + 0.5\, h, \hat{x})$ i otrzymujemy w efekcie punkt D.

Na rysunku widać, że jeśli nachylenie pola wektorowego mocno się zmienia, to pole w punkcie B powinno mieć lepszy kierunek niż w A, czy w E. Jest to oczywiście argument heurystyczny.

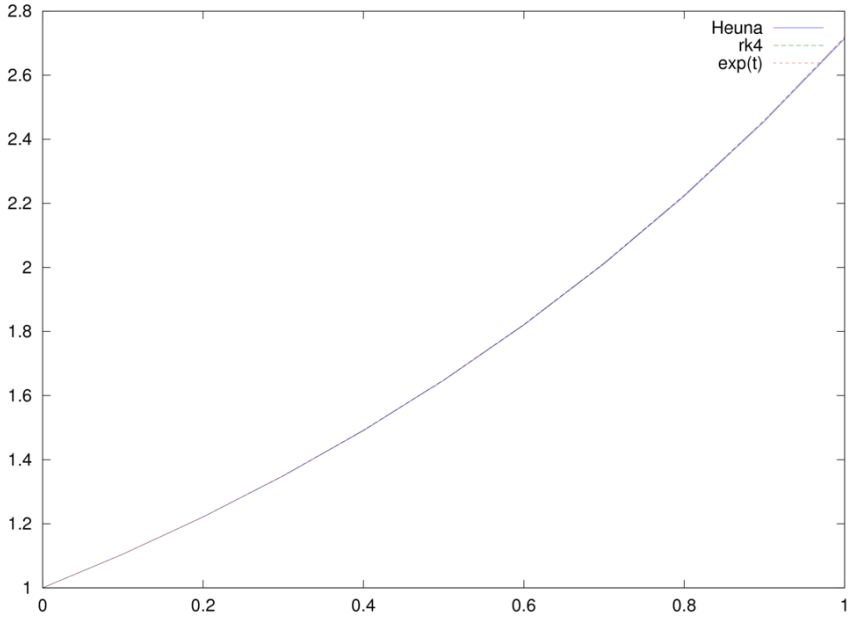


Rysunek 4.7. Schemat Heuna na [0,100] i exp(t). Wykres dla t bliskich 100.

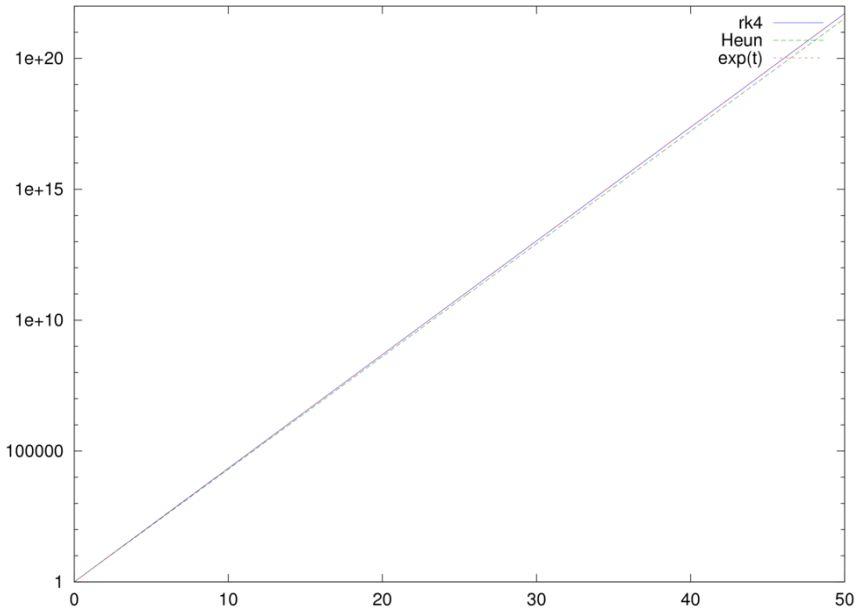


Rysunek 4.8. Graficzne wytłumaczenie zmodyfikowanego schematu Eulera.

Analogicznie konstruuje się schematy Rungego wyższych rzędów poprzez wprowadzenie większej ilości kroków pośrednich, jak również schematy zamknięte Rungego - dopuszczając wartość f_{n+1} w schemacie.



Rysunek 4.9. Schematy Rungego rzędu cztery i schemat Heuna rzędu dwa na $[0, 1]$.



Rysunek 4.10. Schematy Rungego rzędu cztery i schemat Heuna rzędu dwa na $[0, 50]$ w skali pół-logarytmicznej.

Podamy kilka wzorów na powszechnie używany otwarty schemat Rungego-Kutty czwartego rzędu.

Najpierw definiujemy cztery wartości:

$$\begin{aligned} K_1 &= f(t_n, x_n) \\ K_2 &= f(t_n + \tfrac{h}{2}, x_n + \tfrac{h}{2} K_1) \\ K_3 &= f(t_n + \tfrac{h}{2}, x_n + \tfrac{h}{2} K_2) \\ K_4 &= f(t_n + h, x_n + h K_3) \end{aligned} \tag{4.14}$$

i otrzymujemy ostateczny wzór:

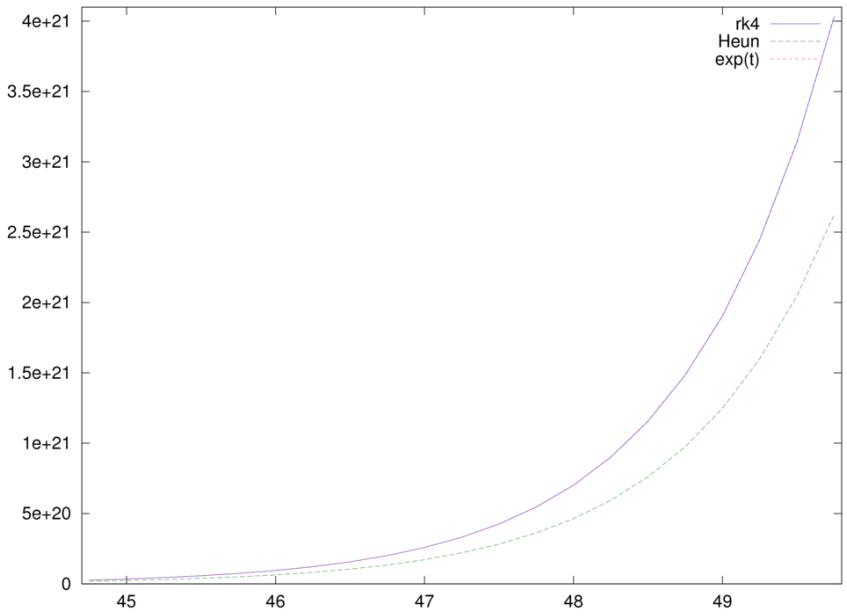
$$x_{n+1} = x_n + \frac{h}{6}(K_1 + 2K_2 + 2K_3 + K_4) \tag{4.15}$$

schematu rzędu cztery (co oczywiście wymaga dowodu).

Istnieje oczywiście cała rodzina otwartych schematów Rungego-Kutty rzędu cztery. Tu podaliśmy tylko przykładowy schemat z tej rodziny.

Popatrzmy jak działa ten schemat w porównaniu ze schematem Heuna dla naszego modelowego zagadnienia $dx/dt = x$ z $x(0) = 1$.

Na odcinku $[0, 1]$ dla $h = 0.1$ oba schematy praktycznie pokrywają się z rozwiązaniem, por. rysunek 4.9. Na rysunku 4.10 widać wykresy rozwiązań na $[0, 50]$ (w skali pół-logarytmicznej),



Rysunek 4.11. Schematy Rungego rzędu cztery i schemat Heuna rzędu dwa na $[44, 50]$.

a na rysunku 4.11 wykres na odcinku $[44, 50]$ - tu już widać, że schemat rzędu cztery jest jednak znacznie lepszy.

W rozdziale 5.4 przedstawimy metodę eksperymentalną badania rzędu schematu. Przy okazji zobaczymy ogromną różnicę w dokładności obu schematów, por. Tabela 5.2.

4.4. Zadania

Ćwiczenie 4.1. Udowodnij, że wzory (4.1) i (4.2) są prawdziwe dla funkcji dostatecznie gładkich.

Wskazówka. Rozwiń funkcje w szereg Taylora.

Ćwiczenie 4.2. Pokaż, że rząd schematów Eulera wynosi jeden, o ile rozwiązanie zadania Cauchy'ego jest klasy C^2.

Ćwiczenie 4.3. Pokaż, że rząd schematu kroku środkowego wynosi k, o ile rozwiązanie zadania Cauchy'ego jest klasy C^{k+1} dla $k = 1, 2$.

Ćwiczenie 4.4. Znajdź rząd schematu Taylora (4.4) dla rozwiązania dostatecznie gładkiego.

Ćwiczenie 4.5. Znajdź wzór na schemat Taylora rzędu trzy.

Ćwiczenie 4.6. Rozpatrzmy rodzinę schematów:

$$x_{n+1} = x_n + h\,(a\,f_n + b\,f_{n+1})$$

Określ rząd schematu w zależności od wartości parametrów a, b. Dla jakich a, b rząd jest największy? Dla jakich wartości parametrów schemat będzie zamknięty, a dla jakich otwarty?

Ćwiczenie 4.7. Wyprowadź otwarty dwukrokowy schemat Adamsa bazujący na wielomianie interpolacyjnym stopnia jeden (żeby policzyć x_{n+2} potrzebujemy h, t_n, x_{n+1}, f_n i f_{n+1}, tak jak opisano w rozdziale 4.1.2). Zbadaj rząd schematu.

Rozwiązanie. Zgodnie z zasadą konstrukcji schematów Adamsa musimy scałkować na odcinku $[t_n, t_n + h]$ wielomian liniowy interpolacyjny $P_1(s) = f_n + \frac{1}{-h}\,(f_n - f_{n-1})(s - t_n)$. Otrzymujemy schemat $x_{n+1} = x_n + 0.5\,h\,(3\,f_n - f_{n-1})$.

Ćwiczenie 4.8. Wyprowadź otwarty trzykrokowy schemat Adamsa bazujący na wielomianie interpolacyjnym stopnia dwa (żeby policzyć x_{n+3} potrzebujemy h, t_n, x_{n+2} i f_n, f_{n+1}, f_{n+2}, tak jak opisano w rozdziale 4.1.2). Zbadaj rząd schematu.

Ćwiczenie 4.9. Wyprowadź zamknięty dwukrokowy schemat Adamsa bazujący na wielomianie interpolacyjnym stopnia dwa (żeby policzyć x_{n+2} potrzebujemy h, t_n, x_{n+1} i f_n, f_{n+1}, f_{n+2}, tak jak opisano w rozdziale 4.1.2). Zbadaj rząd schematu.

Ćwiczenie 4.10 (średnio trudne)**.** Rozpatrzmy rodzinę schematów:

$$x_{n+2} = x_{n+1} + h\,(a\,f_n + b\,f_{n+1} + c\,f_{n+2}).$$

Określ rząd schematu w zależności od wartości parametrów a, b, c. Dla jakich wartości rząd jest największy? Dla jakich wartości parametrów schemat będzie zamknięty, a dla jakich otwarty?

Ćwiczenie 4.11 (trudne)**.** Udowodnij, że schemat (4.15) ma rząd cztery.

Ćwiczenie 4.12 (laboratoryjne)**.** Zbadaj eksperymentalnie metodą połowienia kroków rząd lokalnego błędu schematu dla schematów:

- otwartego schematu Eulera (3.4),
- zamkniętego schematu Eulera (3.5),
- schematu midpoint (4.3),
- schematu Taylora (4.4),
- schematu Heuna (4.13),
- schematu trapezów (4.7),
- zmodyfikowanego schematu Eulera (4.12),
- schematu Rungego rzędu cztery (4.15).

zastosowanych do modelowego zadania $\frac{dx}{dt} = 1 + x^2$ z $x(0) = 0$ z rozwiązaniem $x(t) = \tan(t)$. Tzn. dla ustalonego t, np. $t = 1$ i kolejnych połowionych kroków $h_k = \frac{h_{k-1}}{2} = 2^{-k} h_0$ z $h_0 = 1e - 1$ liczymy lokalny błąd schematu $e_{h_k}(t)$, por. (4.9), i następnie stosunek $\frac{e_{h_k}(t)}{e_{h_{k+1}}(t)}$. Jeśliby ten stosunek wynosił w przybliżeniu 2^{p+1}, to lokalny błąd schematu zachowuje się jak $O(h^{p+1})$, oznacza to, że schemat posiada rząd p przynajmniej dla tego zadania początkowego.

5. Metody dla równań różniczkowych zwyczajnych - teoria zbieżności

W tym rozdziale przedstawimy teorię zbieżności schematów jednokrokowych i wielokrokowych. Rozpatrzymy tylko przypadek skalarny, ale teoria dla układów równań jest analogiczna. Wystarczy tylko moduł zastąpić przez jakąś normę w $\mathbb{R}^m$ np. normę euklidesową.

5.1. Teoria zbieżności schematów jednokrokowych

Teoria zbieżności schematów jednokrokowych jest teorią odrębną od teorii dla schematów wielokrokowych liniowych.

Okaże się, że kluczowym pojęciem jest tu zgodność schematu jednokrokowego - inaczej konsystentność, którą definiujemy następująco:

Definicja 5.1. Schemat jednokrokowy (4.10) jest zgodny (konsystentny), (ang. *consistent*) jeśli:
- ϕ jest ciągłą ze względu na wszystkie zmienne
- $\phi(0,t,x,x) = f(t,x)$ dla wszystkich (t,x).
- ϕ jest lipschitzowska ze względu na zmienne x_n i x_{n+1} tzn. istnieje $L > 0$ takie, że dla wszystkich $x_1, x_2, y_1, y_2 \in U_{x_0}$

$$|\phi(h,t,x_1,x_2) - \phi(h,t,y_1,y_2)| \leqslant L \sum_{k=1}^{2} |x_k - y_k|.$$

Twierdzenie 5.1 (o zbieżności schematów jednokrokowych)**.** *Jeśli rozwiązanie zagadnienia początkowego (3.1) $x \in C^{p+1}([t_0,T])$, schemat jednokrokowy jest zgodny i jest rzędu $p \geqslant 1$, to ten schemat jest zbieżny z rzędem p.*

Dowód. Dowód zostanie tutaj przedstawiony dla prostoty tylko dla schematów otwartych z rzędem p tj. $\phi(h,t,x,y) = \phi(h,t,x)$. Dowód w całej ogólności można znaleźć np. w [23].

Oznaczmy przez $E_n = x_n - x(t_n)$, czyli błąd pomiędzy obliczonym schematem przybliżeniem rozwiązania dla czasu t_n, a dokładną wartością rozwiązania $x(t_n)$. Niech $\tau_n = e_h(t_n) = x(t_{n+1}) - x(t_n) - h\phi(h,t_n,x(t_n))$, czyli τ_n to lokalny błąd schematu dla czasu t_n. Wtedy otrzymujemy, że

$$E_n = E_{n-1} + h(\phi(h,t_{n-1},x_{n-1}) - \phi(h,t_{n-1},x(t_{n-1}))) - \tau_{n-1},$$

a stąd, korzystając ze zgodności schematu, a dokładniej lipschitzowskości funkcji ϕ, por. Definicja 5.1, otrzymujemy

$$|E_n| \leqslant (1 + h * L)|E_{n-1}| + |\tau_{n-1}|.$$

Dalej, poprzez indukcję matematyczną otrzymujemy:

$$|E_n| \leqslant (1 + h * L)^n |E_0| + \sum_{k=0}^{n-1} (1 + h * L)^{n-k-1} |\tau_k|$$

Korzystając z tego, że $(|1+x| \leqslant e^{|x|})$

$$(1+h*L)^n \leqslant e^{n*h*L} \leqslant e^{L*(T-t_0)}$$

dla n takich, że $h*n \leqslant T-t_0$ widzimy, że

$$|E_n| \leqslant e^{L*(T-t_0)}(|E_0| + \sum_{k=0}^{n-1} |\tau_k|)$$

Zauważmy, że $E_0 = 0$. Widzimy też, że

$$|\tau_n| \leqslant e_h.$$

Ponieważ schemat ma rząd p i $x \in C^{p+1}$, to $e_h = O(h^{p+1})$ zatem

$$|E_n| \leqslant e^{L*(T-t_0)} n * e_h \leqslant e^{L*(T-t_0)} \frac{T-t_0}{h} O(h^{p+1}) = O(h^p).$$

□

W szczególności zbieżność z rzędem p oznacza dla ustalonego $t \in [t_0, T]$ i $n*h = t$, że

$$|x_n^h - x(t)| = O(h^p) \to 0 \qquad h \to 0 \quad (n \to \infty).$$

Przykład 5.1. Zgodność otwartego schematu Eulera. W zasadzie sprawdzenie konsystentności (zgodności) schematu jest w tym przypadku oczywiste, ponieważ funkcja $\Phi(h,t,x,y) = f(t,x)$, czyli spełnia założenia zgodności. Dodatkowo wiemy, że schemat ma rząd jeden, więc jeśli x rozwiązanie zadania początkowego (3.1) należy do C^2, to $|x_n - x(t)| = O(h)$ dla $n*h = t-t_0$, co potwierdzają wyniki eksperymentów z tabeli 4.1 w przypadku skalarnym dla równania $\frac{dx}{dt} = x$ z $x(0) = 1$.

5.2. Teoria zbieżności schematów liniowych wielokrokowych

Teoria zbieżności dla schematów wielokrokowych liniowych różni się od teorii zbieżności dla schematów jednokrokowych. Precyzyjniej pisząc - używamy tak samo jak w przypadku schematów jednokrokowych pojęcie rzędu schematu, ale równie ważne jest pojęcie stabilności schematu. Używając nieformalnego języka - stabilność schematu oznacza to, że błędy z poprzednich kroków się nie kumulują, a wręcz zanikają.

5.2.1. Stabilność, zgodność

Na początku rozważmy dowolny schemat liniowy wielokrokowy (4.8) zastosowany do równania z polem wektorowym równym zero tzn. do zagadnienia początkowego:

$$\frac{dx}{dt} = 0 \qquad x(0) = 1,$$

którego jedynym rozwiązaniem jest rozwiązania stałe $x(t) = 1$. Nasz schemat staje się wtedy równaniem różnicowym liniowym jednorodnym o stałych współczynnikach:

$$\alpha_k x_{n+k} + \ldots \alpha_0 x_n = 0. \tag{5.1}$$

Wprowadzając wielomian

$$\rho(\lambda) = \sum_{j=0}^{k} \alpha_j \lambda^j \tag{5.2}$$

otrzymujemy, że jeśli $\xi \neq 0$ jest zerem (pierwiastkiem) wielomianu $\rho(\lambda)$, to ciąg

$$x_n = \xi^n \qquad n = 0, 1, 2, \ldots, \infty$$

jest rozwiązaniem równania różnicowego. Jeśli zero jest dwukrotne , to otrzymujemy nowe rozwiązanie związane z tym pierwiastkiem tzn.:

$$y_n = n * \xi^{n-1}, \qquad n = 0, 1, 2, \ldots, \infty.$$

i tak dalej - jeśli zero jest trzykrotne, to kolejne rozwiązanie:

$$z_n = n * (n-1) * \xi^{n-2}.$$

W każdym razie - co ważne - jeśli jakiś pierwiastek $\rho(\lambda)$ ma moduł $|\xi| > 1$ lub $|\xi| = 1$, ale krotność pierwiastka jest większa od jeden, to istnieją rozwiązania nieograniczone, a dokładnie, zachodzi dla pewnych rozwiązań:

$$|x_n| \to +\infty \qquad n \to \infty.$$

Jeśli pierwiastek $|\xi| < 1$ to zawsze wszystkie rozwiązania z nim związane spełniają

$$|x_n| \to 0 \qquad n \to \infty.$$

Podsumowując: jeśli jakiś pierwiastek ma moduł mniejszy od jeden, albo ma moduł jeden i krotność jeden, to rozwiązania z nim związane są ograniczone. Wydaje się, że wymaganie żeby wszystkie rozwiązania schematu zastosowanego do równania z prawą stroną równą zero były ograniczone jest jak najbardziej uzasadnione.

Dlatego w ten sposób definiujemy stabilność (ang. *stability*) schematu liniowego wielokrokowego (4.8):

Definicja 5.2. Schemat (4.8) jest stabilny, jeśli każdy pierwiastek ξ wielomianu $\rho(\lambda) = \sum_{j=0}^{k} \alpha_j \lambda^j$ spełnia

$$|\xi| \leqslant 1,$$

a w przypadku jeśli $|\xi| = 1$, to krotność pierwiastka ξ wynosi jeden.

Dodatkowo wprowadzamy pojęcie silnej stabilności schematu (ang. *strong stability*):

Definicja 5.3. Schemat (4.8) jest silnie stabilny, jeśli spełnia Definicję 5.2 stabilności, i jeśli ξ jest zerem wielomianu $\rho(\lambda)$ takim, że $|\xi| = 1$, to $\xi = 1$.

Pojęcie silnej stabilności nie wpływa na samą teorię zbieżności schematów, ale można się spodziewać, że schematy silnie stabilne zachowują się lepiej, por. rozdział 5.2.2.

W praktycznych obliczeniach wszystkie używane schematy liniowe wielokrokowe są silnie stabilne.

Zgodność schematu

Możemy oczekiwać, że jednym z rozwiązań naszego równania różnicowego (5.1) będzie rozwiązanie stałe: $x_n = 1$, czyli oczekujemy żeby $\xi = 1$ było pierwiastkiem wielomianu $\rho(\lambda)$, tzn. $\rho(1) = 0$. Ten warunek nazwiemy *prezgodnością* (ang. *preconsistency*) schematu.

Dodatkowo rozpatrzmy drugie proste równanie z warunkiem początkowym:

$$\frac{dx}{dt} = 1 \qquad x(0) = 0,$$

którego rozwiązaniem jest $x(t) = t$. Jeśli zastosujemy schemat do tego równania (zawsze $f_n = 1$), to otrzymujemy następujące liniowe równanie różnicowe o stałych współczynnikach:

$$\sum_{j=0}^{k} \alpha_j x_{n+j}^h - h * \sum_{j=0}^{k} \beta_j = 0 \qquad n \geqslant 0.$$

Dodatkowo wprowadzimy wielomian:

$$\sigma(\lambda) = \sum_{j=0}^{k} \beta_j \lambda^j.$$

Zakładając, że schemat jest dokładny w $t_n = n * h$ dla tego zagadnienia początkowego tzn. wstawiając $x_n = x(n * h) = n * h$ otrzymujemy:

$$\sum_{j=0}^{k} \alpha_j (n + j) - \sum_{j=0}^{k} \beta_j = n * \rho(1) + \rho'(1) - \sigma(1) = 0.$$

Jeśli założymy, że schemat jest prezgodny, to powyższe równanie daje nam:

$$\rho(1) = 0 \qquad \rho'(1) - \sigma(1) = 0. \tag{5.3}$$

Wprowadzamy pojęcie zgodności (konsystentności) schematu liniowego wielokrokowego:

Definicja 5.4. Schemat liniowy wielokrokowy (4.8) jest zgodny (konsystentny), (ang. *consistent*) jeśli zachodzi (5.3).

Nie jest trudno pokazać, że ten warunek jest równoważny temu, że rząd schematu jest co najmniej jeden, tzn. prawdziwe jest następujące stwierdzenie:

Stwierdzenie 5.1. *Schemat liniowy wielokrokowy (4.8) jest zgodny wtedy i tylko wtedy, gdy rząd schematu wynosi co najmniej jeden.*

Dowód pozostawiamy jako zadanie, por. ćwiczenie 5.14.

Zbieżność

Twierdzenie 5.2 (o zbieżności schematów wielokrokowych)**.** *Jeśli rozwiązanie zagadnienia początkowego $x \in C^{p+1}([t_0, T])$, schemat liniowy k-krokowy jest rzędu p dla $p \geqslant 1$ i jest stabilny, oraz wartości startowe x_j^h dla $j = 0, \ldots, k-1$ spełniają nierówność:*

$$\max_{j=0,\ldots,k-1} |x(t_j^h) - x_j^h| \leqslant Ch^p$$

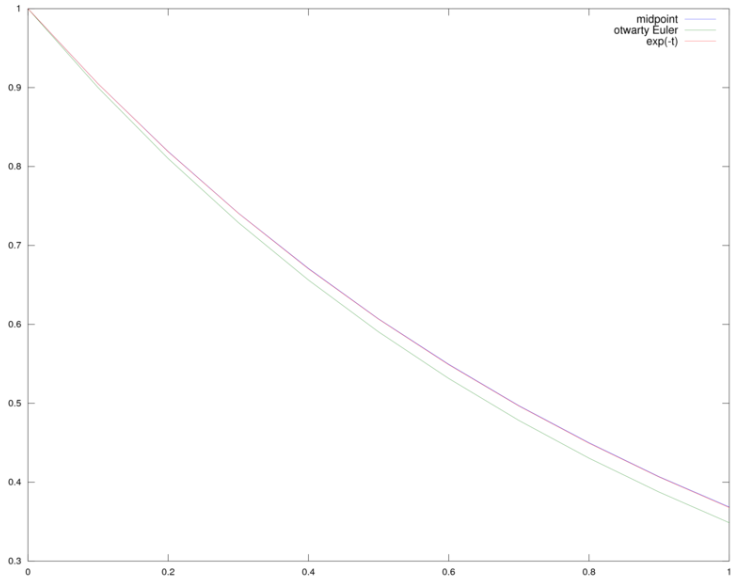
dla pewnej stałej nieujemnej C niezależnej od h, to ten schemat jest zbieżny z rzędem p.

Dowód twierdzenia można znaleźć w [5] lub [23].

Przykład 5.2. Zbieżność otwartego schematu Eulera jako schematu liniowego wielokrokowego. Wiemy już, że rząd otwartego schematu Eulera wynosi jeden. Wystarczy więc zbadać stabilność schematu. Wielomian $\rho(\lambda) = \lambda - 1$ zatem ma jeden pierwiastek $\lambda_1 = 1$ więc schemat jest stabilny, a nawet silnie stabilny. Jeśli rozwiązanie jest klasy C^2, to zbieżność zachodzi z rzędem jeden, co potwierdzają wyniki eksperymentów z tabeli 4.1 w przypadku skalarnym dla równania $\frac{dx}{dt} = x$ z $x(0) = 1$.

5.2.2. Stabilność, a silna stabilność

Rozpatrzmy schemat kroku środkowego (4.3), który jest dwukrokowym schematem liniowym rzędu dwa (zadanie). Nietrudno sprawdzić, że jest to schemat stabilny, ale *nie* silnie stabilny. Rozpatrzmy nasze modelowe równanie liniowe:



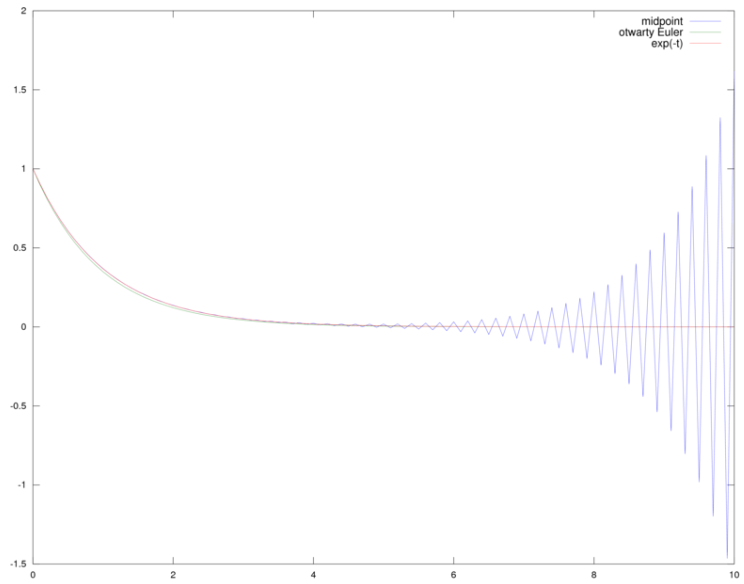
Rysunek 5.1. Schematy midpoint i Eulera otwarty na odcinku [0,1] dla $\frac{dy}{dx} = -y; y(0) = 1$ z $h = 1e - 1$.

$$\frac{dx}{dt} = -x \quad t > 0, \quad x(0) = 1,$$

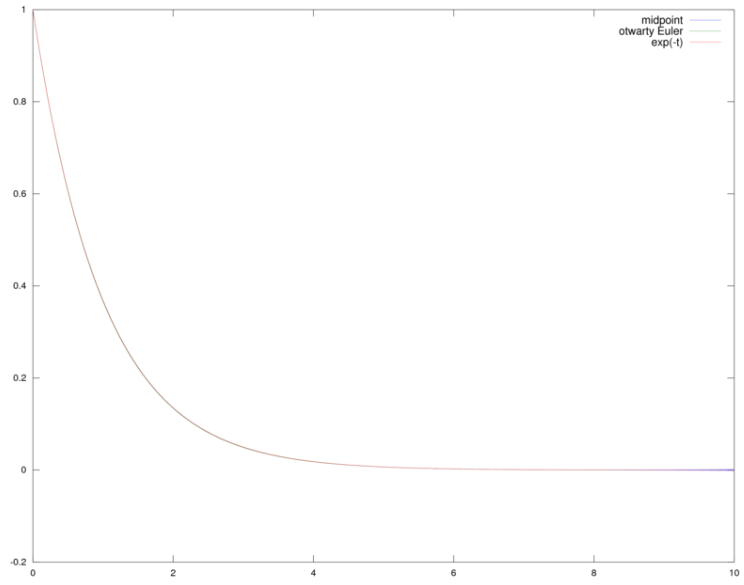
którego rozwiązanie $x(t)$ jest równe $\exp(-t)$. Zastosujmy schemat midpoint biorąc za x_1 dokładną wartość rozwiązania $x_1^h = \exp(-h)$. Narysujmy wykres rozwiązania przybliżonego otrzymanego tym schematem i schematem Eulera otwartym dla $h = 0.1$ na odcinku $[0, 1]$, por. rysunek 5.1, i potem na $[0, 10]$, por. rysunek 5.2. W tym drugim przypadku rozwiązanie przybliżone otrzymane schematem midpoint od pewnego momentu przestaje zachowywać się jak rozwiązanie dokładne $\exp(-t)$, tzn. widzimy coraz większe oscylacje. Weźmy mniejsze $h = 0.01$. Wtedy na $[0, 10]$ wszystko jest w porządku, ale na odcinku $[0, 20]$ znów rozwiązanie otrzymane schematem midpoint zaczyna od pewnego momentu oscylować z coraz większą amplitudą, zamiast maleć do zera. Im h jest mniejsze, tym odcinek na którym rozwiązanie otrzymane schematem midpoint dobrze aproksymuje rozwiązanie równania wyjściowego jest większy, ale zawsze w pewnym momencie pojawi się zaburzenie numeryczne, por. rysunki 5.3 i 5.4.

Numeryczne zaburzenie wynika właśnie z tego, że schemat nie jest silnie stabilny. Można wypisać wzór na rozwiązania równania różnicowego zadanego przez schemat midpoint dla tego równania i wykazać, że po jakimś czasie zawsze pojawią się oscylacje.

Wyniki tu otrzymane nie stoją w jakiekolwiek sprzeczności z teorią zbieżności schematów liniowych wielokrokowych. Można sprawdzić, że schemat midpoint spełnia założenia tej teorii dla tego równania, i że na ustalonym odcinku $[0, T]$ zachodzi zbieżność z rzędem dwa. W praktyce nie należy stosować schematów, które nie są silnie stabilne, ponieważ zawsze mogą pojawić się

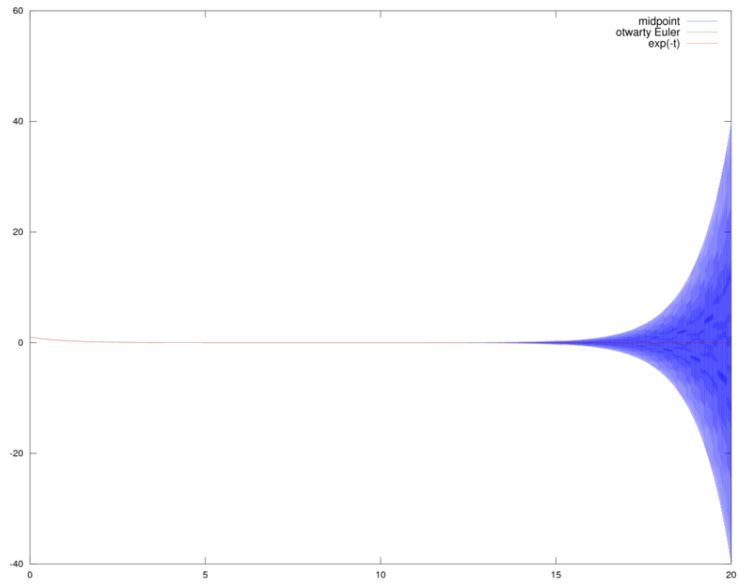


Rysunek 5.2. Schematy midpoint i Eulera otwarty na odcinku [0,10] dla $\frac{dy}{dx} = -y; y(0) = 1$ z $h = 1e - 1$.



Rysunek 5.3. Schematy midpoint i Eulera otwarty na odcinku [0,10] dla $\frac{dy}{dx} = -y; y(0) = 1$ z $h = 1e - 2$.

oscylacje po a priori nie znanym czasie, szczególnie kiedy chcemy rozwiązać równanie na długim odcinku czasu.



Rysunek 5.4. Schematy midpoint i Eulera otwarty na odcinku [0,20] dla $\frac{dy}{dx} = -y; y(0) = 1$ z $h = 1e - 2$.

5.3. Wartości startowe schematów wielokrokowych

W tym miejscu warto zwrócić uwagę, jak w praktycznych obliczeniach implementować schematy liniowe wielokrokowe. Aby przy pomocy danego schematu k-krokowego rzędu p obliczyć x_{n+k}, należy znać poprzednie k przybliżenia $x_{n+k-1}, \ldots, x_n$. Czyli aby wystartować schemat musimy znaleźć odpowiednio dobre (tzn. takie, że $E_j = O(h^p)$, por. Twierdzenie 5.2) startowe przybliżenia $x_0, x_1, \ldots, x_{k-1}$. Wartość x_0 jest zadana jako warunek początkowy, tzn. możemy przyjąć dany warunek początkowy $x_0 = x(t_0)$. Natomiast musimy wyliczyć $x_1, \ldots, x_{k-1}$. W praktyce do wyliczenia tych wartości możemy np. zastosować $k-1$ razy schemat jednokrokowy tego rzędu co najmniej p.

Dla schematu Adamsa-Bashfortha dwukrokowego rzędu dwa za x_1 możemy przyjąć x_1 obliczone jednym krokiem schematu Heuna. Tu warto zauważyć że musimy zastosować tylko jeden krok schematu Heuna. Uwzględniając to widzimy, że x_1 można by też obliczyć stosując dowolny schemat jednokrokowy rzędu jeden np. otwarty schemat Eulera. Ogólniej do obliczenia startowych wartości $x_1, \ldots, x_{k-1}$ dla schematu k-krokowego rzędu p można zastosować schemat jednokrokowy rzędu $p-1$ lub większego.

5.4. Eksperymentalne badanie rzędu zbieżności schematów

Rząd zbieżności schematu dla czasu t można badać eksperymentalnie dla ustalonego zagadnienia początkowego $\frac{dx}{dt} = f(t,x)$ z warunkiem początkowym $x(0) = x_0$ i ze znanym rozwiązaniem $x(t)$ określonym na odcinku $[t_0, T]$. Możemy wtedy przy pomocy naszego schematu, np. schematu Heuna, zmodyfikowanego schematu Eulera i otwartego schematu Eulera obliczać przybliżone wartości x_n^h dla ustalonego $t = t_n^h = t_0 + n * h$, z połowionym krokiem h: $h_0, h_0/2, \ldots, 2^{-k}h_0$ dla ustalonego h_0. Będziemy badać jak zmienia się błąd w kolejnych krokach, a dokładniej, jak zmienia się stosunek błędów dla kolejnych połowionych h.

h	*otwarty Euler*	$\lvert e_n/e_{n+1}\rvert$	*Heun*	$\lvert e_n/e_{n+1}\rvert$	*zmod. Euler*	$\lvert e_n/e_{n+1}\rvert$
$2.5e-01$	$4.61e-02$	0.00	$1.90e-03$	0.00	$6.57e-03$	0.00
$1.1e-01$	$1.98e-02$	2.33	$3.51e-04$	5.40	$1.19e-03$	5.53
$5.3e-02$	$9.23e-03$	2.14	$7.63e-05$	4.60	$2.56e-04$	4.64
$2.6e-02$	$4.47e-03$	2.07	$1.78e-05$	4.28	$5.97e-05$	4.29
$1.3e-02$	$2.20e-03$	2.03	$4.32e-06$	4.13	$1.44e-05$	4.14
$6.3e-03$	$1.09e-03$	2.02	$1.06e-06$	4.07	$3.55e-06$	4.07
$3.1e-03$	$5.44e-04$	2.01	$2.63e-07$	4.03	$8.79e-07$	4.03
$1.6e-03$	$2.71e-04$	2.00	$6.55e-08$	4.02	$2.19e-07$	4.02

Tabela 5.1. Eksperymentalne badanie rzędu zbieżności schematów: otwartego Eulera, Heuna i zmodyfikowanego schematu Eulera poprzez połowienie kroków dla równania $\frac{dx}{dt} = \cos(x)^2$ z $x(0) = 0$ dla $t = 1$, którego rozwiązaniem jest arctan(x). W kolumnach o indeksach parzystych jest podany błąd odpowiedniego schematu dla $t = 1$, a w kolumnach 3, 5 i 7 są stosunki błędów dla kroku $2 * h$ podzielone przez błędy dla kroku h dla odpowiednich schematów.

h	*Heun*	$\lvert e_n/e_{n+1}\rvert$	*rk4*	$\lvert e_n/e_{n+1}\rvert$
$5.3e-01$	$5.96e+03$		$9.09e+01$	
$2.6e-01$	$1.91e+03$	3.12	$6.41e+00$	14.18
$1.3e-01$	$5.29e+02$	3.61	$4.24e-01$	15.12
$6.3e-02$	$1.38e+02$	3.83	$2.73e-02$	15.56
$3.1e-02$	$3.52e+01$	3.92	$1.73e-03$	15.78
$1.6e-02$	$8.88e+00$	3.96	$1.09e-04$	15.89
$7.8e-03$	$2.23e+00$	3.98	$6.81e-06$	15.95
$3.9e-03$	$5.59e-01$	3.99	$4.27e-07$	15.97
$2.0e-03$	$1.40e-01$	4.00	$2.66e-08$	16.02

Tabela 5.2. Eksperymentalne badanie rzędu zbieżności schematów: Heuna i Rungego Kutty rzędu cztery poprzez połowienie kroków dla równania $\frac{dx}{dt} = x$ z $x(0) = 1$ dla $t = 10$. W kolumnach indeksach parzystych jest podany błąd odpowiedniego schematu dla $t = 10$, a w kolumnach 3, i 5 są stosunki błędów dla kroku $2 * h$ podzielone przez błędy dla kroku h dla odpowiednich schematów.

Jeśli błąd dla ustalonego t zachowywałby się jak $O(h^p)$, a dokładniej, gdyby $e_n = |x_n^h - x(t)| = c * h^p + O(h^p)$, to stosunek błędu powinien się zachowywać jak:

$$\frac{\|x_n^{2h} - x(t)\|}{\|x_{2n}^h - x(t)\|} = \frac{c2^p h^p + O(h^{p+1})}{ch^p + O(h^{p+1})} \approx 2^p \qquad h < 1$$

dla dostatecznie małych h, czyli dla schematów Heuna i zmodyfikowanego schematu Eulera jak cztery, a dla otwartego schematu Eulera jak dwa, a dla schematu rzędu cztery jak 2^4 - czyli szesnaście.

W tabeli 5.1 widzimy wyniki eksperymentu dla $t = 1$ dla schematów: otwartego schematu Eulera, Heuna i zmodyfikowanego schematu Eulera zastosowanych do zagadnienia początkowego $\frac{dx}{dt} = \cos(x)^2$ z $x(0) = 0$, którego rozwiązaniem jest $\arctan(x)$, czyli $x(1) = \arctan(1)$.

W tabeli 5.2 przedstawiliśmy wyniki tego samego eksperymentu, ale dla schematu Rungego-Kutty rzędu cztery dla zadania początkowego $\frac{dx}{dt} = x$ z $x(0) = 1$ i dla $t = 10$, czyli z rozwiązaniem

$x(10) = \exp(10)$. Skoro schemat jest rzędu cztery, możemy oczekiwać, że błąd będzie jak $O(h^4)$ dla ustalonego t.

Wyniki uzyskane w tabeli 5.2 potwierdzają nasze przypuszczenie. Dla kroku o połowę mniejszego błąd maleje około $2^4 = 16$ razy dla dostatecznie małych h (im h mniejsze, tym ten stosunek bliższy jest szesnastu). Przy okazji zauważmy, jak ogromna jest różnica w błędzie dla schematu rzędu cztery (Rungego-Kutty), a dla schematu Heuna rzędu dwa. W przypadku tego ostatniego - błąd bezwzględny dla $h = 2 * 10^{-3}$ jest rzędu $1/10$, a dla tego pierwszego błąd jest rzędu 10^{-8}. Z kolei patrząc na błędy względne, w przypadku schematu Heuna błąd jest na poziomie 10^{-5}, a dla schematu Rungego 10^{-12}, czyli poziom błędu w arytmetyce podwójnej precyzji praktycznie jest wystarczająco dokładny.

Błędem bezwzględnym nazywamy e_n, a względnym $e_n/|x(t_n)|$. W przypadku gdy $|x(t_n)|$ jest bardzo duże lub bardzo małe należy rozpatrywać błąd względny choćby z powodu własności arytmetyki zmiennopozycyjnej.

5.5. Zadania

Ćwiczenie 5.1. Zbadaj rząd zbieżności zamkniętego schematu Eulera korzystając z teorii zbieżności dla schematów jednokrokowych.

Ćwiczenie 5.2. Zbadaj rząd zbieżności zamkniętego schematu Eulera korzystając z teorii zbieżności dla schematów wielokrokowych liniowych. Czy schemat jest silnie stabilny?

Ćwiczenie 5.3. Zbadaj rząd zbieżności schematu trapezów dany wzorem (4.7).
1. korzystając z teorii zbieżności dla schematów jednokrokowych,
2. korzystając z teorii zbieżności dla schematów wielokrokowych liniowych.

Czy schemat jest silnie stabilny?

Ćwiczenie 5.4. Zbadaj rząd zbieżności zmodyfikowanego Eulera i schematu Heuna.

Ćwiczenie 5.5. Zbadaj rząd zbieżności schematu punktu środkowego (ang. *midpoint*). Czy schemat jest silnie stabilny?

Ćwiczenie 5.6 (laboratoryjne)**.** Zaimplementuj schematy: otwarty i zamknięty Eulera w octave i przetestuj rzędy zbieżności eksperymentalnie metodą połowionego kroku, jak opisano w rozdziale 5.4 dla równania skalarnego $\frac{dy}{dx} = y^2; y(0) = 1$ oraz dla równania drugiego rzędu $\frac{d^2y}{dx^2} = -y$ z $y(0) = 0, \frac{dy}{dx}(0) = 1$.

Wskazówka. Do rozwiązywania nieliniowego równania przy implementacji schematu zamkniętego w każdym kroku możesz użyć funkcji fsolve().

Ćwiczenie 5.7 (laboratoryjne)**.** Zaimplementuj schemat trapezów, por. (4.7), w octave i przetestuj rząd zbieżności eksperymentalnie metodą połowionego kroku, jak opisano w rozdziale 5.4 dla równania skalarnego $\frac{dy}{dx} = -y; y(0) = 1$ oraz dla równania drugiego rzędu $\frac{d^2y}{dx^2} = -y$ z $y(0) = 0, \frac{dy}{dx}(0) = 1$.

Wskazówka. Do rozwiązywania nieliniowego równania w każdym kroku możesz użyć funkcji fsolve().

Ćwiczenie 5.8 (laboratoryjne)**.** Zaimplementuj zmodyfikowany schemat Eulera oraz schemat Heuna w octave i przetestuj rzędy zbieżności eksperymentalnie metodą połowionego kroku, jak opisano w rozdziale 5.4.

Ćwiczenie 5.9 (laboratoryjne)**.** Zaimplementuj schemat Rungego rzędu cztery, dany wzorem (4.15) w octave i przetestuj rząd zbieżności schematu eksperymentalnie metodą połowionego kroku, jak opisano w rozdziale 5.4.

Ćwiczenie 5.10 (laboratoryjne)**.** Zaimplementuj schemat (ang. *midpoint*) w octave i przetestuj rząd zbieżności schematu eksperymentalnie metodą połowionego kroku, jak opisano w rozdziale 5.4, przyjmując, że $x_1 = y(t_0 + h)$, czyli jest równe dokładnemu rozwiązaniu. Dla równania $\frac{dy}{dx} = -y$ z $y(0) = 1$ zastosuj schemat na długim odcinku czasu dla ustalonego h. Czy rozwiązanie zachowuje się tak jak tego oczekujemy? Zmniejsz h i zobacz czy sytuacja się poprawia?

Ćwiczenie 5.11. Zbadaj rząd, stabilność i rząd zbieżności otwartego dwukrokowego schematu Adamsa, por. Ćwiczenie 4.7.

Ćwiczenie 5.12 (laboratoryjne)**.** Zaimplementuj w octave otwarty dwukrokowy schemat Adamsa, por. Ćwiczenie 4.7. Następnie przetestuj eksperymentalnie rząd zbieżności tego schematu metodą połowionego kroku, jak opisano w rozdziale 5.4 dla $\frac{dy}{dt} = -y$ z $y(0) = -1$. Biorąc za x_1:

1. $x_1 = y(h) = e^{-h}$ czyli rozwiązanie dokładne dla $t = h$.
2. x_1 obliczone schematem Heuna czyli schematem Rungego-Kutty rzędu dwa.
3. x_1 obliczone schematem Eulera otwartym czyli schematem rzędu jeden.

Ćwiczenie 5.13. Znajdź dokładne rozwiązanie $(y_k)_{k=1}^{\infty}$ schematu różnicowego, który jest schematem punktu środkowego zastosowanym dla równania $\frac{dy}{dx} = -y$. Pokaż, że jeśli $y_0 = 1$, a y_1 różne od jednej konkretnej wartości (co np. odpowiada warunkom startowym $y_0 = 1, y_1 = \exp(-h)$), to $\lim_{n\to\infty} |y_n| = +\infty$, czyli pojawią się niepotrzebne kumulujące się zaburzenia numeryczne.

Ćwiczenie 5.14 (trudne)**.** Udowodnij stwierdzenie 5.1.

Wskazówka. Z tego, że rząd schematu jest co najmniej jeden, wynika od razu warunek zgodności schematu (wystarczy rozważyć oba równania $\frac{dy}{dx} = 0$ z $y(0) = 1$ i $\frac{dy}{dx} = 0$ z $y(0) = 0$). Natomiast aby pokazać, że ze zgodności schematu wynika to, że rząd schematu jest większy od jeden, należy zastosować wzór Taylora.

Ćwiczenie 5.15. Czy schemat Adamsa (por. rozdział 4.1.2) może nie być stabilny, ewentualnie nie być silnie stabilny? Uzasadnij podając ewentualnie kontrprzykład niestabilnego (czy nie będącego silnie stabilnym) schematu Adamsa.

6. Sztywność, zmienny krok całkowania i metoda strzałów

W tym rozdziale zajmiemy się ważnymi schematami rozwiązywania tzw. sztywnych układów równań różniczkowych zwyczajnych. Omówimy schematy ze zmiennym krokiem całkowania i metodę strzałów rozwiązywania zadań brzegowych.

6.1. Sztywne równania różniczkowe zwyczajne

Dość trudno jest podać precyzyjnie poprawną matematycznie definicję sztywności dla dowolnego zadania różniczkowego zwyczajnego. My przyjmiemy definicję pragmatyczną za [13]:

Definicja 6.1. Równanie różniczkowe zwyczajne nazywamy sztywnym (ang. *stiff*), jeśli numeryczne schematy zamknięte, w szczególności metody zamknięte Adamsa, działają zdecydowanie lepiej niż schematy otwarte przybliżonego rozwiązywania zagadnień początkowych.

Oczywiście definicja nie jest do końca precyzyjna. Podamy też inne definicje sztywności (ang. *stiffness*) np. dla zagadnień liniowych, jakkolwiek powyższa definicja jest dla nas wygodna, ponieważ podkreśla to, że równania sztywne rozwiązujemy przy pomocy schematów zamkniętych. Za chwilę podamy kilka przykładów sztywnych równań różniczkowych, aby przekonać się, że pojawiają się one dość często w realistycznych modelach nauk przyrodniczych, por. rozdział 6.2.

6.1.1. Przypadek skalarny

W rozdziale 3.3.1 już zobaczyliśmy, że dla modelowego zadania skalarnego:

$$\frac{dx}{dt} = a * x \quad x(0) = 1 \qquad a < 0.$$

rozwiązanie uzyskane przy pomocy otwartego schematu Eulera zachowuje własności rozwiązania $x(t) = \exp(a*t)$, tzn. jest dodatnie i malejące do zera dla $t \to +\infty$ tylko wtedy, gdy jest spełniony warunek: $h < -1/a$. W przypadku gdy $|a|$ jest bardzo duże, warunek ten wymusza, że musimy stosować bardzo małe h. Natomiast rozwiązanie uzyskane przy pomocy zamkniętego schematu Eulera zachowuje własności powyższe rozwiązania dla dowolnego $h > 0$.

Schemat zamknięty Eulera można zatem uznać za lepszy od schematu otwartego dla tego zagadnienia dla ujemnego a o dużym module.

6.1.2. Przypadek wielowymiarowy

Załóżmy, że rozpatrujemy jednorodne liniowe równanie różniczkowe zwyczajne ze stałymi współczynnikami:

$$\frac{dx}{dt} = A * x \qquad x(t_0) = \vec{\beta},$$

gdzie A - to macierz o stałych współczynnikach taka, że w pewnej bazie jest ona diagonalizowalna, tzn. istnieje macierz nieosobliwa C taka, że

$$A = C\Lambda C^{-1}$$

z

$$\Lambda = \begin{pmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_N \end{pmatrix}$$

Załóżmy, że wszystkie $\lambda_k < 0$, wtedy oczywiście rozwiązaniem jest

$$x(t) = Ce^{\Lambda(t-t_0)}C^{-1}\vec{\beta}.$$

Oznaczmy k-tą kolumnę macierzy C przez $\vec{c}_k$, tzn. $C = (\vec{c}_1, \ldots, \vec{c}_N)$, wtedy otrzymujemy:

$$x(t) = \sum_k \alpha_k \vec{c}_k e^{\lambda_k(t-t_0)}.$$

dla $\vec{\alpha} = (\alpha_1, \ldots, \alpha_N)^T = C^{-1}\vec{\beta}$. Jeśli zastosujemy otwarty schemat Eulera do tego zagadnienia, to analogicznie jak w poprzednim rozdziale otrzymujemy, że

$$x_n \quad = \quad (I + h * A)x_{n-1} = C(I + h * \Lambda)C^{-1}x_{n-1} = C(I + h * \Lambda)^n\vec{\alpha} = \sum_k \alpha_k \vec{c}_k(1 + h\lambda_k)^n.$$

Czyli: o ile $|\lambda_k| >> 1$ ($\lambda_k < 0$) tym odpowiednia składowa rozwiązania $\vec{c}_k e^{\lambda_k(t-t_0)}$ szybciej dąży do zera. Z drugiej strony warunek na to, aby odpowiadająca składowa rozwiązania dyskretnego otrzymanego otwartym schematem Eulera nie zmieniała znaku i zbiegała do zera wynosi:

$$h < 1/|\lambda_k|,$$

czyli jest to warunek ograniczający dopuszczalny zakres wartości h.

Widzimy zatem, że otwarty schemat Eulera dla takiego równania jest zupełnie niepraktyczny na dłuższych odcinkach czasu, ponieważ ograniczenie na h związane jest ze składowymi rozwiązań, które najszybciej zanikają, czyli na dłuższym odcinku czasu nie mają większego wpływu na rozwiązanie. Z kolei dla schematu zamkniętego Eulera nie otrzymujemy żadnych ograniczeń na h, ponieważ:

$$x_n = C(I - h * \Lambda)^{-n}\vec{\alpha} = \sum_k \alpha_k(1 - h\lambda_k)^{-n}.$$

Widzimy, że dla tego typu równań schemat otwarty zachowuje się gorzej niż odpowiedni schemat zamknięty, co jest zgodne z naszą oryginalną definicją sztywności. W ogólnym przypadku, gdy wszystkie części rzeczywiste wartości własnych macierzy A są ujemne sytuacja jest analogiczna.

Zatem definiujemy zadanie liniowe jednorodne jako sztywne, jeśli:

1. $Re\,\sigma(A) \subset \{x : x < 0\}$
2. $\frac{\max_{\lambda_k \in \sigma(A)} |Re\,\lambda_k|}{\min_{\lambda_k \in \sigma(A)} |Re\,\lambda_k|}$ jest duże.

Tutaj $\sigma(A)$ oznacza zbiór wartości własnych macierzy A. Oczywiście w tej definicji nie jest doprecyzowane co oznacza «duże», ale łatwo określić, że jeśli stosunek w drugim punkcie jest równy dziesięć, to układ nie jest sztywny, a jeśli 10^{20} to układ jest sztywny. Rozszerza się powyższą definicję sztywności na układy równań nieliniowych przyjmując, że układ:

$$\frac{dx}{dt} = F(t, x)$$

jest sztywny w obszarze G i dla t z odcinka (a, b) jeśli Jakobian F, czyli $D_xF(t, x)$ spełnia powyższą definicję dla każdego $x \in G$ i $t \in (a, b)$.

Wadą powyższej definicji sztywności jest to, że nie obejmuje np. układu skalarnego $\dot{x} = a * x$ dla $a < 0$.

6.2. Przykłady schematów sztywnych

W tym rozdziale podamy kilka przykładów równań sztywnych, por. [13].

6.2.1. Oscylator Van der Pola

Równanie Van der Pola opisujące oscylator z nieliniowym tłumieniem:

$$\frac{d^2y}{dx^2} - a * (1 - y * y) * \frac{dy}{dx} + y = 0 \qquad a > 0,$$

gdzie $a > 0$ jest parametrem, dla dużego $a > 1$ np. $a = 1000$ powyższe równanie jest sztywne.

6.2.2. Reakcje chemiczne

Rozważmy następujące reakcje chemiczne, które symbolicznie opiszemy następująco:

$$\begin{aligned} A &\overset{0.04}{\rightarrow} B && \text{(wolna)} \\ B + B &\overset{3*10^7}{\rightarrow} C + B && \text{(bardzo szybka)} \\ B + C &\overset{10^4}{\rightarrow} A + C && \text{(szybka)} \end{aligned}$$

co prowadzi do następującego układu równań różniczkowych zwyczajnych:

$$\begin{array}{lrcrcrcl} A: & x_1' & = & -0.04 * x_1 & + & 10^4 x_2 * x_3 & & x_1(0) = 1 \\ B: & x_2' & = & 0.04 * x_1 & - & 10^4 x_2 * x_3 & -3 * 10^7 x_2^2 & x_2(0) = 1 \\ C: & x_3' & = & & & & 3 * 10^7 x_2^2 & x_3(0) = 1 \end{array}$$

Czytelnikowi pozostawiamy sprawdzenie z pomocą octave'a, że np. schematy otwarte nie działają najlepiej dla tego problemu.

6.2.3. Równania paraboliczne

Ten przykład jest szczególnym przypadkiem dyskretyzacji równań, których metody dyskretyzacji omawiane są później dokładniej w rozdziale 14.

Rozpatrzmy równanie paraboliczne:

$$\frac{\partial u}{\partial t}(t,x) = \frac{\partial^2 u}{\partial x^2}(t,x) + f(t,x) \qquad x \in (0,1) \quad t \in (0,T]$$

z warunkami brzegowymi $u(t,0) = u(t,1) = 0$ i początkowym $u(0,x) = u_0(x)$.

Dyskretyzując je względem zmiennej przestrzennej x metodą różnic skończonych, tzn. wprowadzając siatkę $x_k = k * h$ dla $k = 0, \ldots, N$ z $h = 1/N$ i przybliżając drugą pochodną przez

$$\frac{\partial^2 u}{\partial x^2}(t,x) \approx \frac{u(t,x-h) - 2 * u(t,x) + u(t,x+h)}{h^2},$$

otrzymujemy następujący układ równań różniczkowych zwyczajnych:

$$\frac{du_k}{dt}(t) = \frac{u_{k-1}(t) - 2 * u_k(t) + u_{k+1}(t)}{h^2} + f_k(t) \qquad k = 1, \ldots, N-1$$

z $u_0(t) = u_N(t) = 0$ (warunki brzegowe) i warunkiem początkowym $u_k(0) = u_0(k * h)$ i $f_k(t) = f(k * h, t)$, por. rozdział 14.

Oczekujemy, że $u_k(t) \approx u(t, x_k)$ gdzie $u(t,x)$ jest rozwiązaniem wyjściowego problemu. Nietrudno zauważyć, że powyższy układ równań zwyczajnych można zapisać jako

$$\vec{u}' = -h^{-2} A_N \vec{u} + \vec{f}, \qquad \vec{u}(0) = (u_0(x_1), \ldots, u_0(x_{N-1}))^T$$

dla $\vec{u} = (u_1(t), \ldots, u_{N-1}(t))^T$, $\vec{f}(t) = (f_1(t), \ldots, f_{N-1}(t))^T$ i

$$A_N = \begin{pmatrix} 2 & -1 & 0 & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & 0 & -1 & 2 \end{pmatrix} \tag{6.1}$$

Można pokazać, że wartości własne A_N są dodatnie i

$$\frac{\max_{\lambda \in \sigma(A_N)} \lambda}{\min_{\lambda \in \sigma(A_N)} \lambda} = O(N^2),$$

czyli układ jest sztywny dla dużych N, czyli małych h.

Dla przykładowych dużych N możemy sprawdzić, czy rzeczywiście tak jest w octave, że stosunek największej do najmniejszej wartości własnej wynosi ok. 4000 dla $N = 100$, a dla $N = 10^3$ ten stosunek wynosi ok. $4 * 10^5$. Tu podajemy kod octave obliczający ten stosunek:

```
N=100;
A=diag(2*ones(N,1))-diag(ones(N-1,1),1)-diag(ones(N-1,1),-1);
max(eig(A))/min(eig(A))
#ans = 4133.6
#a teraz dla N=10^3 - co zajmie dluzsza chwile
N=10^3;
A=sparse(diag(2*ones(N,1)))-sparse(diag(ones(N-1,1),1))-sparse(diag(ones(N-1,1),-1));
ev=eig(A);
max(ev)/min(ev)
#ans = 4.0610e+05
```

6.3. Schematy zamknięte. Predyktor-korektor

Schematy zamknięte stosujemy dla zadań sztywnych.

Aby obliczyć kolejne przybliżenie x_n w schematach zamkniętych, musimy rozwiązać liniowy lub nieliniowy układ równań np. dla zamkniętego schematu Eulera:

$$x_n = x_{n-1} + h * f_n,$$

czyli w każdym kroku musimy rozwiązać układ równań:

$$g(x_n) = 0$$

dla funkcji $g(x) = x - h * f(t_n, x) - x_{n-1}$.

W ogólności dla zamkniętego schematu k-krokowego liniowego musimy w każdym kroku rozwiązać układ równań względem x_n:

$$0 = g(x_n) = \alpha_k x_n - h * \beta_k f(t_n, x_n) + \sum_{j=0}^{k-1} \alpha_j x_{n+j-k} - h \sum_{j=0}^{k-1} \beta_j f_{n+j-k}.$$

Analogiczna sytuację widzimy dla zamkniętych schematów jednokrokowych. Powyższe równanie (układ równań) możemy rozwiązać przy pomocy różnych metod np. metody Newtona, czy jakiejś wersji metody iteracji prostej, zob. np. [18], [17]. Zauważmy, że w przypadku otwartego schematu Eulera x_n jest punktem stałym dla funkcji $G_n(x) = h*f(t_n, x) - x_{n-1}$, czyli naturalne jest zastosowanie następującej metody iteracyjnej: dla danego x^0 liczymy

$$x^k = h * f(t_n, x^k) + x_{n-1} = G_n(x^k) \qquad k = 1, \ldots.$$

Z odpowiedniej gładkości pola wektorowego f wynika, że $x \mapsto f(t_n, x)$ jest funkcją lipschizowską względem x (lokalnie na $\overline{K} = \overline{K}(x_{n-1}, \epsilon)$) ze stałą Lipschitza L_f i f jest ograniczona na kuli $\overline{K}$, tzn. dla $x \in \overline{K}$ zachodzi $\|f(t_n, x)\| \leqslant M$. Wtedy

$$\forall x \in \overline{K} \qquad \|G_n(x) - x_{n-1}\| \leqslant h * \|f(t_n, x)\| \leqslant h * M,$$

i stała Lipschitza G_n wynosi $h * L_f$, ponieważ

$$\forall x, y \in \overline{K} \qquad \|G_n(x) - G_n(y)\| = \|h * f(t_n, x) - h * f(t_n, y)\| \leqslant h * L_f \|x - y\|.$$

Zatem jeśli dla odpowiednio małego h zachodzi $h*M < \epsilon$ i $h*L_f \leqslant \alpha < 1$, to G_n jest kontrakcją na $\overline{K}$. Z tego wynika istnienie x_n i zbieżność metody iteracji prostej. W przypadku innych schematów zamkniętych możemy skonstruować analogiczne wersje metody iteracji prostych.

Postawmy kwestię - jak dobierać startowe przybliżenie x^0.

Pierwsza opcja to: wziąć $x^0 = x_{n-1}$. Z ciągłości rozwiązania wynika, że jeśli h jest dostatecznie małe, to $x_n \approx x(t_n)$ i $x_{n-1} \approx x(t_n - h)$ są sobie bliskie, a dokładnie zachodzi $\|x(t_n - h) - x_n\| \approx O(h)$.

Zastanówmy się, czy można dobrać x^0 lepiej?

Istnieje możliwość, żeby za x^0 brać przybliżenie $x(t_n)$ obliczone jednym krokiem otwartego schematu tego samego rzędu p co schemat zamknięty (oczywiście zbieżnym z tym samym rzędem) tzn.

$$x^0 = \hat{x}_n = \Phi(h, t_n, x_{n-l} \ldots, x_{n-1})$$

gdzie $x_{n-1}, \ldots, x_{n-l}$ są obliczone wczesniej naszym schematem zamkniętym, a $\hat{x}_n = \Phi(h, t_n, \hat{x}_{n-l} \ldots, \hat{x}_{n-1})$ jest dowolnym l- krokowym otwartym schematem zbieżnym z rzędem p, por. (4.5).

Wtedy widzimy, że $\|x^0 - x(t_n)\| \approx O(h^p)$ więc i o ile h dostatecznie małe $\|x^0 - x_n\| \approx O(h^p)$.

W takim przypadku schemat otwarty nazywamy *predyktorem*, a schemat zamknięty, który naprawdę stosujemy do rozwiązania zadania początkowego - *korektorem.* Podsumowując; nazwy schemat *predyktor-korektor* używa się względem schematu zamkniętego rzędu p, zaimplementowanego w ten sposób, że kolejne x_n^h przybliżenie $x(t_n^h)$ obliczone jest poprzez zastosowanie w każdym kroku czasowym jakiejś metody iteracyjnej rozwiązywania nieliniowego równania (układu równań) z przybliżeniem startowym obliczonym odpowiednim pojedyńczym krokiem schematu otwartego tego samego rzędu (predyktorem). Metoda iteracyjna niekoniecznie musi być taka, jak opisana powyżej. Do rozwiązywania nieliniowego układu równań można stosować też np. metodę Newtona, czy jeszcze inną metodę iteracyjną, por. np. [18] lub [25].

W praktyce bierze się odpowiednie pary schematów tego samego rzędu: np. otwarty schemat Eulera za predyktor i zamknięty schemat Eulera za korektor, czy ogólniej - schemat otwarty Adamsa-Bashfordsa rzędu k za predyktor ze schematem zamkniętym Adamsa-Moultona rzędu k jako korektorem. Popatrzmy, jak wygląda przykładowa implementacja schematu predyktor-korektor w przypadku schematów Eulera: otwartego schematu Eulera wziętego jako predyktor i zamkniętego schematu Eulera, który tu pełni rolę korektora dla równania $\dot{x} = 1 + x * x$ z $x(0) = 0$ z rozwiązaniem $x(t) = \tan(t)$. Zaimplementowaliśmy powyższy schemat w octave biorąc jako metodę iteracyjnego rozwiązywania równania nieliniowego w każdym kroku funkcję octave `fsolve()`:

```
function [X,t]=predkoreuler(f,t0=0,x0=1,N=100,h=1.0/N)
# Parametry funkcji:
#f − wskaznik do pola wektorowego − funkcji dwóch argumentów f(x,t)
# przy czym x0 − wektor pionowy dlugosc M;
# przyklad definicji wskaznika do prostego pola wekt.: f=@(x,t) −x;
#t0 − czas początkowy
#h − stały krok dla schematu Eulera
#N − ilość kroku schematu
#Funkcja zwraca macierz X wymiaru (N+1)xM długości N+1 taka ze
#X(k,:) jest przybliżeniem rozwiazania w punkcie czasu t0+(k−1)*h
#oraz wektor t dlugosci N+1 z dyskretnymi punktami czasowymi
global xx hh tt
hh=h;
M=length(x0);
X=zeros(N+1,M);
t=zeros(N+1,1);
xx=X(1,:)=x0;
tt=t(1)=t0;
for k=2:N+1,
  xp=xx+h*f(xx,tt); #predyktor
  g=@(x) x − hh*f(x,tt) − xx; #funkcja pomocnicza dla zamkniętego schematu Eulera
  X(k,:)=xx=fsolve(g,xp); #rozwiązujemy równanie − korektor
  tt+=hh;
  t(k)=tt;
endfor
endfunction
```

6.4. Adaptacyjny dobór kroku całkowania

Stałe w twierdzeniach o zbieżności schematów są znacznie zawyżone i dobór kroku całkowania w oparciu o szacowania z tych twierdzeń jest niepraktyczny. Czy można jakoś oszacować błąd na bieżąco i zmieniać krok całkowania w zależności od tych oszacowań?

Załóżmy, że chcemy użyć konkretnego schematu jednokrokowego rzędu k przybliżonego rozwiązywania zadania początkowego (3.1) takiego, że przy założeniu odpowiedniej gładkości pola wektorowego f otrzymujemy, że błąd metody spełnia dla $0 < h < 1$:

$$e_n^h = x_n^h - x(t_n^h) = e(t_n^h)h^k + O(h^{k+1}).$$

dla $t_n^h = t_0 + n * h \in [a, b]$, x rozwiązania (3.1), x_n^h rozwiązania przybliżonego obliczonego naszym schematem i pewnej funkcji $e(t)$. Można pokazać, że tak rzeczywiście jest, i że funkcja $e(t)$ jest rozwiązaniem odpowiedniego równania różniczkowego. Najważniejsze zaś jest to, że $e(t)$ nie zależy od h. Oznaczmy przez $x(t; h) := x_n^h$ rozwiązanie otrzymane przy pomocy schematu dla ustalonego h i dla $t = t_n^h$, a przez $e(t; h) := e_n^h$ oznaczmy błąd metody w punkcie $t = t_n^h = t_0 + n * h$.

Wtedy

$$e(t; h) = e(t)h^k + O(h^{k+1}),$$
$$e(t; h/2) = e(t)\frac{h^k}{2^k} + O(h^{k+1}).$$

Postępując podobnie jak w ekstrapolacji Richardsona, jeśli odejmiemy stronami te równości to otrzymamy:

$$e(t;h) - e(t;h/2) = x(t;h) - x(t;h/2) = e(t)\frac{h^k}{2^k}(2^k - 1) + O(h^{k+1}),$$

czyli otrzymujemy:

$$E(x;h/2) := \frac{x(t;h) - x(t;h/2)}{2^k - 1} = e(t)\frac{h^k}{2^k} + O(h^{k+1}) \approx e(t;h/2).$$

dla $h << 1$, a zatem:

$$\|e(t)\| \approx \|E(x;h/2)\| * \frac{2^k}{h^k}.$$

Otrzymaliśmy w ten sposób estymator błędu.

Jeśli chcemy otrzymać błąd na poziomie ϵ i dla pewnego t obliczyliśmy:

$$\|E(x;h/2)\| \approx \|e(t)\|\frac{h^k}{2^k} \approx \|e(t;h/2)\|,$$

to możemy wyliczyć h_1, dla którego błąd będzie na poziomie ϵ, tzn. przyjmując

$$\|e(t;h_1)\| \approx \|e(t)\|h_1^k = \epsilon$$

otrzymujemy

$$h_1 = \left(\frac{\epsilon}{\|e(t)\|}\right)^{1/k} \approx \frac{h}{2}\left(\frac{\epsilon}{\|E(x;h/2)\|}\right)^{1/k}. \tag{6.2}$$

Następnie możemy zastosować ten schemat z krokiem h_1.

Oczywiście adapatacyjną zmianę kroku całkowania (ang. *adaptive step control*) można stosować i do zwiększania kroku w celu obniżania kosztu obliczeń.

To znaczy, że jeśli $\|E(x;h/2)\| > \epsilon$, to możemy zmniejszyć krok zgodnie z powyższym wzorem i wtedy powtarzamy obliczenia z mniejszym krokiem h_1. Jeśli $\|E(x;h/2)\| \leqslant \epsilon$ to za przybliżenie $x(t)$ weźmiemy $x(t;h/2)$, a do następnego kroku możemy przyjąć nowy większy krok h_1 z (6.2).

Oczywiście zamiast połowienia kroku możemy obliczać $x(t;h/q)$ dla $q = 3$ lub 4 i wtedy otrzymujemy analogiczne wzory.

Można też, zamiast stosowania tego samego schematu dwa razy z krokiem h i potem $h/2$, obliczać przybliżenie $x(t)$, schematem rzędu k, a potem większego rzędu np. $k+1$, jak to się dzieje np. w metodzie Rungego-Fehlberga, gdzie stosuje się schematy Rungego-Kutty czwartego rzędu i Rungego-Kutty piątego rzędu, por. rozdział 17.2 w [24].

6.5. Metoda strzałów

Metoda strzałów (ang. *shooting method*) służy rozwiązywaniu zadań brzegowych. Rozpatrujemy w tym przypadku równanie różniczkowe zwyczajne, w którym część warunków początkowych zastępujemy liniowymi lub nieliniowymi warunkami brzegowymi, tzn. szukamy funkcji klasy C^1 na odcinku $[a,b]$ spełniającej:

$$\begin{aligned} \frac{dx}{dt} &= f(t,x), \qquad t \in (a,b) \\ g(x(a),x(b)) &= 0 \end{aligned} \tag{6.3}$$

dla $g : U \subset \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}^{2m}$ danej funkcji co najmniej ciągłej.

Proszę zauważyć, że ogólnie takie zadanie nie musi mieć rozwiązania nawet w prostym przypadku np.

$$\frac{d^2x}{dt^2} = -x \qquad x(0) = 0 \quad x(\pi) = 1.$$

Rozwiązanie ogólne tego równania to $c_1 \sin(t) + c_2 \cos(t)$ i z powyższych warunków brzegowych otrzymujemy sprzeczne warunki na c_2: $c_1 \sin(0) + c_2 \cos(0) = c_2 = 0$ i $c_1 \sin(\pi) + c_2 \cos(\pi) = -c_2 = 1$.

Jeśli istnieje rozwiązanie zadania brzegowego (6.3), to oczywiście jest to szczególny przypadek rozwiązania zadania początkowego (dla pewnej wartości s):

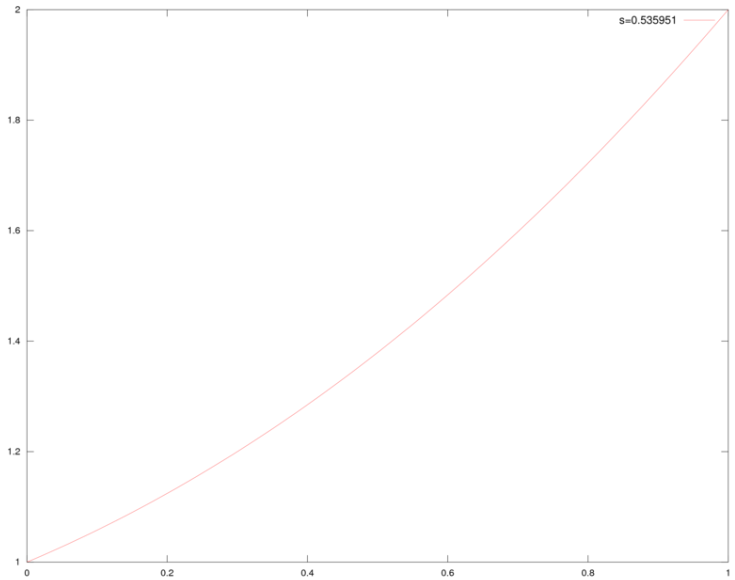
$$\begin{aligned} \frac{d^2x}{dt^2} &= f(t,x) \qquad t \in (a,b), \\ x(a) &= s. \end{aligned} \tag{6.4}$$

Dodatkowo wiemy, że jeśli f jest funkcją ciągłą, to wartość rozwiązania powyższego zadania początkowego dla $t = b$, tzn. $x(b;s)$ jest funkcją ciągłą względem s. A jeśli f jest klasy C^k, to $x(b;s)$ ma taką samą gładkość jak f, por. np. [23].

Jeśli istnieje rozwiązanie (6.3), to dla pewnego s^* zachodzi $g(s, x(b;s^*)) = 0$. Sprowadziliśmy zadanie brzegowe do zadania nieliniowego znalezienia pierwiastka układu:

$$F(s) := g(s; x(b;s)) = 0.$$

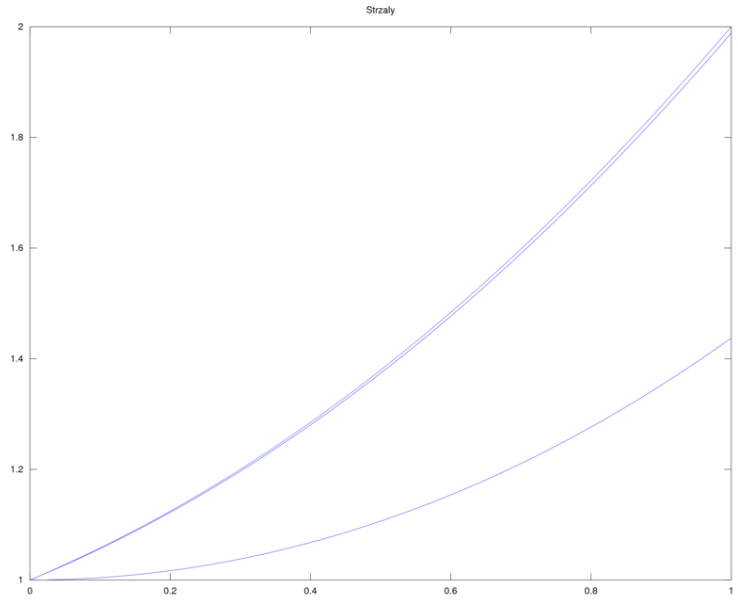
Do rozwiązania tego układu możemy zastosować jakąś metodę rozwiązywania układów równań



Rysunek 6.1. Metoda strzałów - przybliżone rozwiązanie zadania brzegowego: $\frac{d^2y}{dx^2} = \sin(y)$ z $y(0) = 1$ i $y(1) = 2$.

nieliniowych, np. metodę bisekcji (o ile zadanie jest skalarne), czy metodę Newtona lub iteracji prostych, por. np. [18].

Można się spytać: jak obliczyć wartość $F(s)$ dla danego s. Trzeba obliczyć $x(b;s)$, które jest wartością rozwiązania zadania początkowego (6.4) dla $t = b$ z warunkiem początkowym $x(a) = s$.



Rysunek 6.2. Kilka strzałów tzn. wykresów przybliżonych rozwiązań zadania początkowego: $\frac{d^2y_k}{dx^2} = \sin(y_k)$ z $y_k(0) = 1$ i $\frac{dy_k}{dx}(0) = s_k$ dla kolejnych iteracji s_k.

Zwykle nie znamy rozwiązań ogólnych tego równania, więc musimy zastosować jakiś schemat rozwiązywania zadania początkowego.

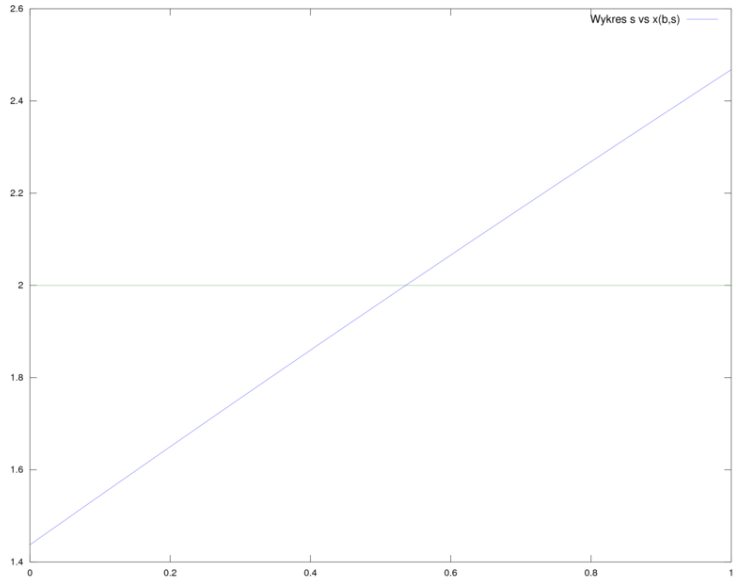
Dla przykładu zastosowaliśmy metodę strzałów do rozwiązania zadania:

$$\frac{d^2y}{dx^2} = \sin(y) \qquad y(0) = 1 \quad y(1) = 2.$$

Wykorzystaliśmy metodę rozwiązywania równań zwyczajnych w octave lsode(), w połączeniu z funkcją octave'a rozwiązywania równań nieliniowych fsolve().

Otrzymaliśmy, że dla $s = 0.53595$ błąd wynosi w przybliżeniu 10^{-8}. Na rysunku 6.1 widzimy wykres rozwiązania, a na rysunku 6.2 widać wykresy przybliżeń rozwiązania, tzn. rozwiązania zadania początkowego z $\dot{x}= s_k$ dla s_k wartości kolejnych iteracji metody Newtona. Na rysunku 6.3 widzimy wykres funkcji $F(s)$.

Niestety metoda strzałów w wielu przypadkach może być bardzo niestabilna. Rozpatrzmy bardzo proste liniowe zadanie: $\frac{d^2y}{dx^2} + y(x) = 0$ z warunkami brzegowymi $y(0) = y(20) = 1$, dla którego znamy rozwiązanie $y(t) = (e^{t-20} + e^{-t})/(1 + e^{-20})$. Zastosowanie metody strzałów z wykorzystaniem standardowej metody rozwiązywania równań zwyczajnych octave'a, czyli funkcją lsode(), daje rozwiązanie przybliżone, dla którego błąd w $t = 20$ wynosi ok. 190. Wynika to z tego, że wartość rozwiązania zadania początkowego $\frac{d^2y}{dx^2} + y(x) = 0$ $y(0) = 1$ $y'(0) = s$ dla $t = 20$, tzn. $y(20; s)$, jest bardzo niestabilna. Małe zaburzenie s powoduje ogromną zmianę wyniku. W tym przypadku możemy funkcję $y(t; s)$, wyliczyć analitycznie, co pozostawiamy jako zadanie. Natomiast zastosowanie metody różnic skończonych daje dobre wyniki, por. rozdział 7.



Rysunek 6.3. Wykres funkcji $s \mapsto y(1;s)$ dla y rozwiązania zadania początkowego: $\frac{d^2y}{dx^2} = \sin(y)$ z $y(0) = 1$ i $\frac{dy}{dx}(0) = s$.

6.6. Zadania

Ćwiczenie 6.1. Rozpatrzmy następujące zadanie brzegowe:

$$\frac{d^2x}{dt^2}(t) - a(t)x(t) = f(t), \quad x(0) = \alpha, \quad x(b) = \beta.$$

dla f funkcji C^∞.

1. Pokaż, że to zadanie ma jednoznaczne rozwiązanie dla stałego współczynnika $a = Const \geqslant 0$.
2. Przy założeniu, że współczynnik a jest stały, wyznacz wszystkie wartości a, dla których powyższe zadanie może nie mieć rozwiązania.
3. Pokaż, że jeśli znamy rozwiązania zadania początkowego x_1 i x_2:

$$\frac{d^2x}{dt^2}(t) - a(t)x(t) = f(t) \quad x(0) = \alpha \quad x'(0) = s_k$$

dla różnych $s_1 \neq s_2$ takie, że $x_1(b) \neq x_2(b)$, to możemy wyznaczyć wzór na s od $s_1, s_2, x_1(b), x_2(b)$ takie, że rozwiązanie zadania początkowego dla tego równania z warunkiem początkowym $x(0) = \alpha \quad x'(b) = s$ będzie rozwiązaniem wyjściowego zadania brzegowego.

Ćwiczenie 6.2 (laboratoryjne)**.** Rozpatrzmy następujące zadanie brzegowe:

$$\frac{d^2x}{dt^2}(t) - a(t)x(t) = f(t), \quad x(0) = \alpha, \quad x(b) = \beta$$

dla $a(t) \geqslant 0$.

Zaimplementuj w octave metodę rozwiązywania zadania brzegowego metodą strzałów korzystając z rozwiązania poprzedniego zadania. W szczególności przetestuj dla $a(t) = 1$ i $f(t) = 0$ z warunkami brzegowymi $x(0) = x(b) = 1$ dla $b = 1, 10, 20$. Porównaj wynik z rozwiązaniem dokładnym $x(t) = (e^{t-b} + e^{-t})/(1 + e^{-b})$.

Wskazówka. Rozwiąż na odcinku $[0, b]$ *korzystając z funkcji octave* lsode() *zadanie początkowe dla tego równania z dwoma różnymi warunkami początkowym* $x_k(0) = \alpha$ *i* $x'_k(0) = s_k$ *dla* $k = 1, 2$ *i* $s_1 \neq s_2$*. Następnie oblicz* s *takie, że rozwiązanie zadania początkowego z* $x(0) = \alpha$ *i* $x'(0) = s$ *będzie rozwiązaniem wyjściowego zadania brzegowego.*

Ćwiczenie 6.3 (laboratoryjne)**.** Bazując na otwartym schemacie Eulera zaimplementuj w octave schemat z adaptacyjnym krokiem całkowania korzystający ze wzoru (6.2) w rozdziale 6.4. Następnie dla równania $y' = -y$ z $y(0) = 1$ sprawdź błąd tego schematu dla $t = 1$ i $t = 100$.

Ćwiczenie 6.4 (częściowo laboratoryjne)**.** Wyprowadź wzór analogiczny do (6.2) w rozdziale 6.4 dla schematu drugiego rzędu obliczając $x(t; h/q)$ dla $q = 3$, tzn. wzór na oszacowanie błędu bazujący na przybliżeniach rozwiązania otrzymanych danym schematem dla h i $h/3$. Zastosuj otrzymane wzory dla zadania początkowego z poprzedniego zadania i schematu Heuna.

Ćwiczenie 6.5. Udowodnij, że macierz A_N dana wzorem (6.1) jest symetryczna, nieosobliwa i nieujemnie określona, czyli dodatnio określona.

Wskazówka. Nieujemną określoność najprościej udowodnić z twierdzenia Gerszgorina, por. [17]. A nieosobliwość macierzy - wprost zakładając, że istnieje niezerowy wektor w jądrze macierzy i dochodząc do sprzeczności.

Ćwiczenie 6.6 (laboratoryjne)**.** Zaimplementuj w octave schemat predyktor-korektor biorąc za korektor schemat trapezów rzędu dwa, a za predyktor schemat Heuna. Do rozwiązywania nieliniowego układu równań zastosuj funkcję octave'a fsolve(). Przetestuj rząd takiego schematu metodą połowionego kroku, jak opisano w rozdziale 5.4, dla równania $\dot{x} = 1 + x * x$ z $x(0) = 0$ dla $t = 1$ z rozwiązaniem $x(t) = \tan(t)$ i dla równania wahadła porównując z rozwiązaniem otrzymanym dla równania wahadła przy pomocy funkcji octave'a lsode().

Ćwiczenie 6.7 (laboratoryjne)**.** Zaimplementuj w octave schemat predyktor-korektor biorąc za korektor zamknięty schemat Eulera, a za predyktor otwarty schemat Eulera. Do rozwiązywania nieliniowego układu równań zastosuj swoje metody rozwiązywania równań nieliniowych tzn.:

1. metodę iteracji prostych, jak opisano w rozdziale 6.3,
2. wielowymiarową metodę Newtona.

Przetestuj rząd takiego schematu metodą połowionego kroku, jak opisano w rozdziale 5.4, dla równania $\dot{x} = 1 + x * x$ z $x(0) = 0$ z rozwiązaniem $x(t) = \tan(t)$ dla $t = 1$ i dla równania wahadła porównując z rozwiązaniem otrzymanym dla równania wahadła przy pomocy funkcji octave'a lsode(). Porównaj czas i ilość iteracji potrzebne do wyliczenia x_n każdą z tych metod przy tym samym warunku stopu metody.

Wskazówka. Metoda Newtona rozwiązywania $G(x) = 0$ *jest zdefiniowana następująco:* $x^{k+1} = x^k + h^k$ *dla* $DG(x^k)h^k = -G(x^k)$*. Układ równań liniowych* $Ay = b$ *możemy rozwiązać w octave przy użyciu operatora backslash tzn.* y=A \ b.

7. Metoda różnic skończonych dla równań eliptycznych drugiego rzędu

W tym rozdziale przedstawimy idee metody różnic skończonych na dwóch modelowych przykładach. Ze wszystkich metod przybliżonego rozwiązywania równań różniczkowych cząstkowych metoda ta wydaje się najbardziej intuicyjna w konstrukcji. W klasycznym sformułowaniu danego równania różniczkowego zamiast pochodnych rozpatrujemy ich przybliżenia. Na zadanej siatce rozpatrujemy przybliżenia pochodnych za pomocą różnic skończonych, czyli ilorazów różnicowych.

7.1. Modelowe zadanie jednowymiarowe

Rozpatrzmy następujące zagadnienie brzegowe:

$$\begin{aligned} Lu(x) = -u''(x) + c * u(x) &= f(x) \qquad x \in \Omega = (a,b), \\ u(a) &= \alpha, \\ u(b) &= \beta \end{aligned} \tag{7.1}$$

dla nieujemnej stałej c, ustalonego odcinka $[a,b]$ i znanych wartości α, β.

Na podstawie tego modelowego zadania opiszemy ideę metody różnic skończonych (MRS).

Przyjmijmy następujące oznaczenia na różnicę skończoną w przód (ang. *forward finite difference*) i różnicę skończoną w tył (ang. *backward finite difference*) :

$$\begin{aligned} \partial_h u(x) &= h^{-1}(u(x+h) - u(x)), \\ \overline{\partial}_h u(x) &= h^{-1}(u(x) - u(x-h)). \end{aligned} \tag{7.2}$$

dla $h > 0$. Będziemy często opuszczali dolny indeks h, jeśli h będzie ustalone.

Dla ustalonego kroku $h > 0$ rozważmy następującą aproksymację drugiej pochodnej:

$$\partial\overline{\partial}_h u(x) = \partial\overline{\partial} u(x) = \overline{\partial}\partial u(x) = \frac{u(x-h) - 2u(x) + u(x+h)}{h^2}. \tag{7.3}$$

Nietrudno zauważyć, że jeśli funkcja u jest klasy C^4 w otoczeniu x to:

$$u''(x) = \partial\overline{\partial} u(x) + O(h^2). \tag{7.4}$$

co pozostawiamy jako zadanie, zob. ćwiczenie 7.1.

Wprowadzając siatkę (ang. *mesh*), czyli zbiór dyskretny dla $h = (b-a)/N$:

$$\overline{\Omega}_h = \{a + k * h\}_{k=0}^{N}$$

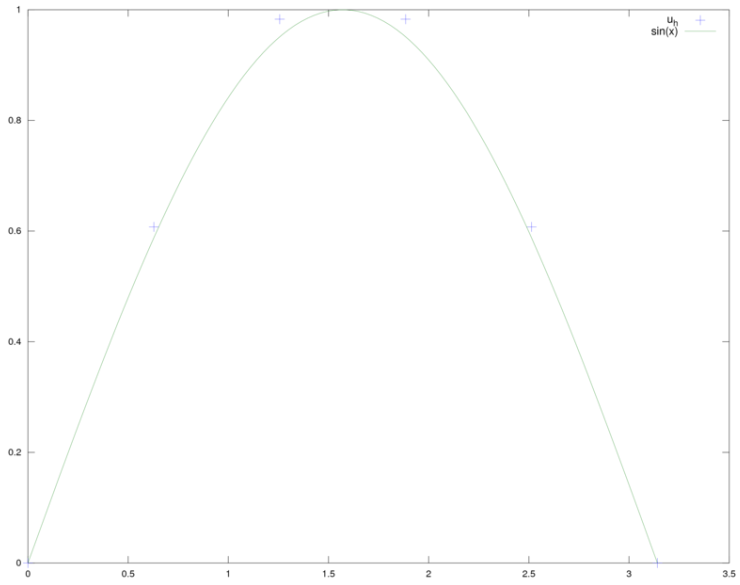
z $\Omega_h = \{a + k * h\}_{k=1}^{N-1}$ i $\partial\Omega_h = \{a + k * h\}_{k=0,N} = \{a, b\}$, możemy zdefiniować następujące zadanie dyskretne: znaleźć u_h funkcję określoną na siatce $\overline{\Omega}_h$ taką, że

$$-\partial\overline{\partial} u_h(x) + c * u_h(x) = f(x) \qquad x \in \Omega_h \tag{7.5}$$

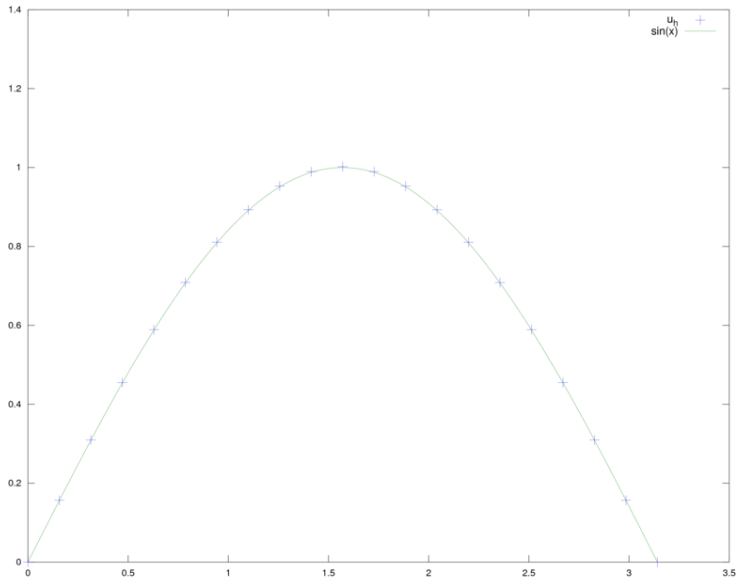
$$u(x) = g(x) \qquad x \in \partial\Omega_h \tag{7.6}$$

Przyjmujemy, że $g : \partial\Omega \to \mathbb{R}$ przyjmuje wartości $g(a) = \alpha$ i $g(b) = \beta$.

Możemy przypuszczać, że u_h będzie aproksymowało w jakimś sensie u rozwiązanie zadania wyjściowego w punktach siatki. Jak porównać u_h określone na zbiorze dyskretnym z u



Rysunek 7.1. Wykres rozwiązania dokładnego $u(x) = \sin(x)$ zadania (7.1) i rozwiązania przybliżonego (7.5) oznaczonego plusami dla siatki sześciopunktowej .



Rysunek 7.2. Wykres rozwiązania dokładnego $u(x) = \sin(x)$ zadania (7.1) i rozwiązania przybliżonego (7.5) oznaczonego plusami dla dwudziestu punktów.

określonym na całym domknięciu obszaru? Możemy porównać te funkcje w punktach siatki licząc błąd:

$$\|u_h - r_h u\|_{\infty,h} = \max_{x \in \overline{\Omega}_h} |u_h(x) - u(x)|$$

dla $r_h u$ funkcji obcięcia (ang. *restriction*) określonej na siatce, przyjmującej wartości u w punktach siatki, ale możemy też badać dyskretną normę Euklidesową L_h^2, czyli pierwiastek z sumy kwadratów błędów w punktach siatki przeskalowanych przez h:

$$\|u_h - r_h u\|_{0,h} = \sqrt{h \sum_{k=0}^{N} |u_h(x_k) - u(x_k)|^2}$$

dla $x_k = a + k * h$. Popatrzmy na wykres rozwiązania zadania dyskretnego dla $c = 0, a = 0, b = \pi$ i $f(x) = \sin(x)$, dla którego rozwiązaniem jest $u(x) = \sin(x)$ dla sześciu punktów i dla dwudziestu punktów, por. rysunki 7.1 i 7.2. Widać, że dla dwudziestu punktów siatki rozwiązanie przybliżone pokrywa się z rozwiązaniem dokładnym w punktach siatki.

Będziemy badać błąd w normach $\|\cdot\|_{\infty,h}$ i $\|\cdot\|_{0,h}$ dla połowionych kroków, tzn. dla $2^{-k}h_0$ dla ustalonego $h_0 = \pi/10$. Wyniki w tabeli 7.1 sugerują, że błędy dyskretne w obu normach są rzędu dwa, tzn. że $\|r_h u - u_h\|_{\infty,h} = O(h^2)$ i $\|r_h u - u_h\|_{0,h} = O(h^2)$. W kolejnych rozdziałach wyjaśnimy dlaczego tak jest. Jak się okaże wyniki eksperymentu są zgodne z oszacowaniami otrzymanymi teoretycznie.

N	$\|e_h\|_{\infty,h}$	$\|e_h\|_{\infty,h}/\|e_{2h}\|_{\infty,2h}$	$\|e_h\|_{0,h}$	$\|e_h\|_{0,h}/\|e_{2h}\|_{0,2h}$
10	$8.265e-03$		$1.036e-02$	
20	$2.059e-03$	$4.01e+00$	$2.580e-03$	$4.01e+00$
40	$5.142e-04$	$4.00e+00$	$6.445e-04$	$4.00e+00$
80	$1.285e-04$	$4.00e+00$	$1.611e-04$	$4.00e+00$
160	$3.213e-05$	$4.00e+00$	$4.027e-05$	$4.00e+00$
320	$8.032e-06$	$4.00e+00$	$1.007e-05$	$4.00e+00$
640	$2.008e-06$	$4.00e+00$	$2.517e-06$	$4.00e+00$
1280	$5.020e-07$	$4.00e+00$	$6.292e-07$	$4.00e+00$
2560	$1.255e-07$	$4.00e+00$	$1.573e-07$	$4.00e+00$
5120	$3.138e-08$	$4.00e+00$	$3.933e-08$	$4.00e+00$

Tabela 7.1. Badanie błędu dyskretnego dla dyskretyzacji różnicami skończonymi zadania $-u'' = \sin(x)$ dla $x \in [0, \pi]$ z $u(0) = u(\pi) = 0$ dla którego znamy rozwiązanie $u(x) = \sin(x)$. W kolumnie drugiej podajemy normę błędu $e_h = r_h u - u_h$ w dyskretnej normie maksimum, a w kolumnie czwartej w dyskretnej normie L^2. W kolumnach trzeciej i piątej podajemy stosunek odpowiedniej normy błędu dla danego h względem normy błędu dla $2 * h$.

Przyjmując oznaczenie $u_k = u_h(x_k) = u_h(a + k * h)$ otrzymujemy następujący układ równań liniowych:

$$\begin{aligned} u_0 &= g(a), \\ \tfrac{1}{h^2}(-u_{k-1} + 2 * u_k - u_{k-1}) + c * u_k &= f(x_k) = f_k \qquad k = 1, \ldots, N-1, \\ u_N &= g(b). \end{aligned}$$

Wstawiając $u_0 = g(a)$ i $u_N = g(b)$ do układu równań otrzymujemy następujący układ równań:

$$\frac{1}{h^2}\begin{pmatrix} 2+c*h^2 & -1 & 0 & \cdots & 0 \\ -1 & 2+c*h^2 & -1 & \vdots & \\ 0 & \ddots & \ddots & \ddots & \vdots \\ & & -1 & 2+c*h^2 & -1 \\ 0 & \cdots & 0 & -1 & 2+c*h^2 \end{pmatrix}\begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_{N-2} \\ u_{N-1} \end{pmatrix} = \begin{pmatrix} f_1 + \frac{1}{h^2}g(a) \\ u_2 \\ \vdots \\ u_{N-2} \\ u_{N-1} + \frac{1}{h^2}g(b) \end{pmatrix}, \tag{7.7}$$

czyli układ z macierzą trójdiagonalną, który można rozwiązać np. metodą *przeganiania* (wersja rozkładu LU dla macierzy trójdiagonalnej) kosztem $O(N)$ dla $N \leqslant 10^6$, czy nawet większych N (w zależności od dostępnego komputera).

7.2. Modelowe zadanie dwuwymiarowe

W tym rozdziale rozpatrzymy modelowe zadanie eliptyczne dwuwymiarowe na kwadracie jednostkowym $\overline{\Omega} = [0,1]^2$. Zadanie polega na znalezieniu $u \in C^2(\Omega) \cap C(\overline{\Omega})$ takiego, że

$$\begin{aligned} Lu(x) = -\triangle u(x) + c * u(x) &= f(x) \qquad x \in \Omega \\ u(s) &= g(s) \qquad s \in \partial\Omega, \end{aligned} \tag{7.8}$$

gdzie $\triangle u = \sum_{k=1}^{2} \frac{\partial^2 u}{\partial x_k^2}$, c jest ustaloną nieujemną stałą, f - to funkcja ciągła na Ω, a g - to funkcja ciągła na $\partial\Omega$. Zakładamy, że istnieje jednoznaczne rozwiązanie (7.8).

Dla ustalonego kroku $h > 0$ rozważmy następującą aproksymację drugiej pochodnej cząstkowej (por. (7.3)):

$$\partial\overline{\partial}_{k,h}u(x) = \partial\overline{\partial}_k u(x) = \overline{\partial}\partial_k u(x) = \frac{u(x - h\vec{e}_k) - 2u(x) + u(x + h\vec{e}_k)}{h^2}. \qquad k = 1,2, \tag{7.9}$$

gdzie $\vec{e}_k$ jest k-tym wersorem.

Wprowadzamy siatkę (zbiór dyskretny) dla $h = 1/N$:

$$\overline{\Omega}_h = \{(k*h, l*h)\}_{k,l=0}^{N} \tag{7.10}$$

i jej odpowiednie podzbiory $\Omega_h = \{(k*h, l*h)\}_{k,l=1}^{N-1} \subset \Omega$ i $\partial\Omega_h = \{(k*h, l*h)\}_{k,l=0,N} \subset \partial\Omega$.

Definiujemy następujące zadanie dyskretne: chcemy znaleźć u_h funkcję określoną na siatce $\overline{\Omega}_h$ taką, że

$$\begin{cases} -\sum_{k=1,2} \partial\overline{\partial}_k u_h(x) + c * u_h(x) = f(x) & x \in \Omega_h, \\ u(x) = g(x) & x \in \partial\Omega_h \end{cases} \tag{7.11}$$

Tak, jak w przypadku jednowymiarowym, możemy porównywać błąd w punktach siatki w dyskretnej normie maksimum:

$$\|u_h - r_h u\|_{\infty,h} = \max_{x \in \overline{\Omega}_h} |u_h(x) - u(x)|$$

dla $r_h u$ zdefiniowanego analogicznie, tzn. funkcji określonej na siatce przyjmującej wartości u w punktach siatki lub w dyskretnej normie L_h^2:

$$\|u_h - r_h u\|_{0,h} = \sqrt{h^2 \sum_{k,l=0}^{N} |u_h(x_{k,l}) - u(x_{k,l})|^2}$$

N	$\|z_h\|_{\infty,h}$	$\|z_h\|_{\infty,h}/\|z_h\|_{\infty,2h}$	$\|z_h\|_{0,h}$	$\|z_h\|_{0,h}/\|z_h\|_{0,2h}$
10	$5.211e-05$		$2.789e-05$	
20	$1.317e-05$	$3.96e+00$	$7.011e-06$	$3.98e+00$
40	$3.298e-06$	$3.99e+00$	$1.755e-06$	$3.99e+00$
80	$8.254e-07$	$4.00e+00$	$4.389e-07$	$4.00e+00$
160	$2.064e-07$	$4.00e+00$	$1.097e-07$	$4.00e+00$

Tabela 7.2. Badanie błędu dyskretnego dla dyskretyzacji różnicami skończonymi dwuwymiarowego zadania $-\triangle u = 2 * \sin(x) * \cos(x)$ na $x \in (0,1)^2$, dla którego znamy rozwiązanie $u(x) = \sin(x)\cos(x)$ z odpowiednim warunkiem brzegowym Dirichleta na $\partial\Omega$. W kolumnie drugiej podajemy normę błędu $z_h = r_h u - u_h$ w dyskretnej normie maksimum, a w kolumnie czwartej w dyskretnej normie L^2. W kolumnach trzeciej i piątej podajemy stosunek odpowiednich norm dla danego h względem normy błędu dla $2 * h$.

dla $x_{k,l} = (k*h, l*h)$. W Tabeli 7.2 podane są wyniki obliczeń dla dyskretyzacji (7.11) zadania (7.8) z $c = 0$ ze znanym rozwiązaniem $u(x) = \sin(x)\cos(x)$ i odpowiednio dobranymi $f(x,y) = 2\sin(x)\cos(x)$ i $g(x,y) = u(x,y)$ dla x na brzegu kwadratu. Stosunki norm dyskretnych dla danego h względem błędu dla $2 * h$ sugerują, że w tym przypadku widzimy rząd zbieżności kwadratowej w obu normach, tzn. że $\|r_h u - u_h\|_{\infty,h} = O(h^2)$ i $\|r_h u - u_h\|_{0,h} = O(h^2)$, tak samo jak w przypadku jednowymiarowym. Oczywiście obliczenia czyli rozwiązywanie odpowiedniego układu równań liniowych jest teraz bardziej kosztowne, jako że np. dla $N = 160$ czyli $h = 1/160$, otrzymujemy układ równań liniowych z $M = 159^2 = 25281$ niewiadomymi i z macierzą o $M^2 = 6.3912 * 10^8$ elementach. Jednak macierz jest pasmowa - o paśmie szerokości $N-1$ i o około $4 * M$, czyli ma 10^5 niezerowych elementów. Zatem możemy jeszcze zastosować specjalne, bezpośrednie metody rozwiązywania układów równań liniowych, takie jak odpowiednia wersja rozkładu LU, por. np. [9], czy - jeśli trzymamy tę macierz w formacie rzadkim, np. spakowanych kolumn, czy wierszy - to możemy zastosować jakąś bezpośrednią metodę dla macierzy rzadkich, np. metodę frontalną, por. [8]. Warto zauważyć, że w tym szczególnym przypadku obszaru $\Omega = (0,1)^2$ znamy wartości i wektory własne tej macierzy i możemy zastosować specjalne metody rozwiązywania tego układu równań z wykorzystaniem algorytmu szybkiej transformaty Fouriera (FFT) (ang. *Fast Fourier Transform*), por. [10].

Moglibyśmy też rozważyć siatkę $\overline{\Omega}_h$ z różnymi rozmiarami h_k $k = 1, 2$ względem obu osi tzn.:

$$\overline{\Omega}_h = \{(k * h_1, l * h_2)\}_{k=0,\ldots,N; l=0,\ldots,M}$$

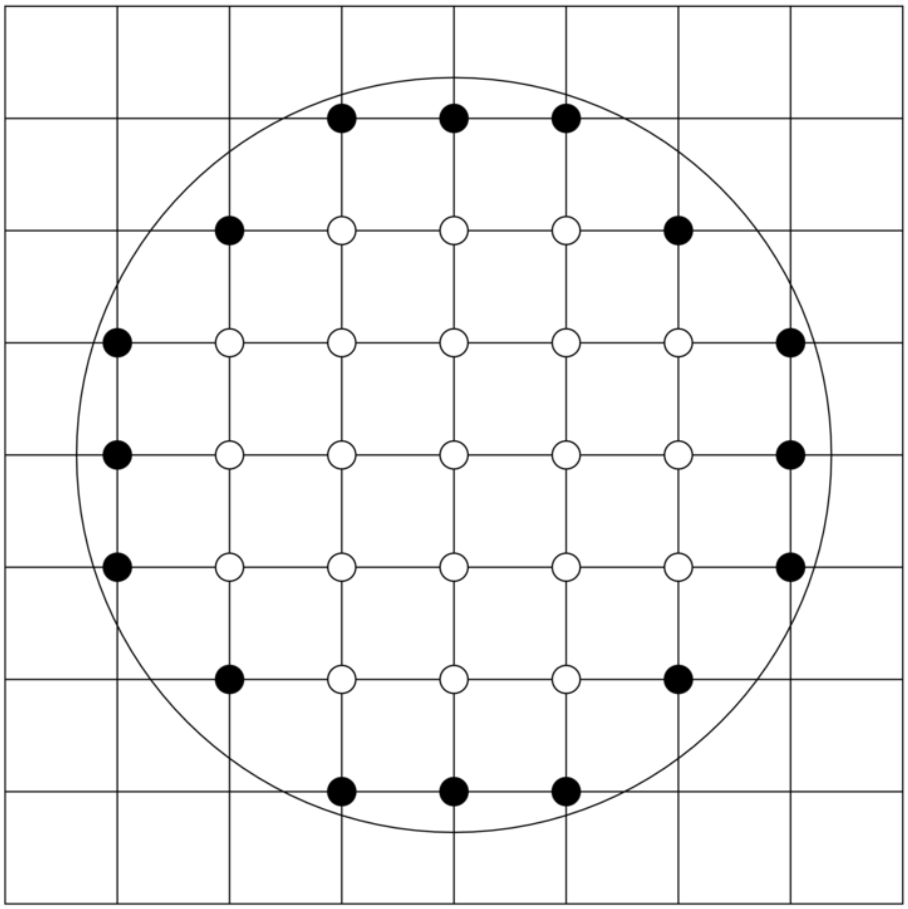
dla $h_1 = 1/N$ i $h_2 = 1/M$ i jej odpowiednie podzbiory $\Omega_h = \overline{\Omega}_h \cap \Omega$ i $\partial\Omega_h = \overline{\Omega}_h \cap \partial\Omega$.

Wtedy oczywiście otrzymalibyśmy trochę inny układ równań, ale generalnie o tych samych właściwościach.

7.2.1. Warunki brzegowe dla obszaru o skomplikowanej geometrii

Zauważmy, że dla obszaru $\overline{\Omega} = [0,1]^2$ możemy tak dobrać kroki siatki, aby otrzymać w sposób naturalny punkty siatki leżące na $\partial\Omega$. W przypadku obszarów o bardziej skomplikowanej geometrii, które nie są prostokątami, czy sumami prostokątów, często nie ma takiej możliwości. Taka sytuacja ma miejsce np. gdy Ω jest kołem.

Dla dowolnego Ω i naszej aproksymacji różnicowej Laplasjanu wprowadzamy pomocniczą siatkę na $\mathbb{R}^2$ postaci $S_h = \vec{x}_0 + (k * h_1, l * h_2)_{k,l \in \mathbb{Z}}$ dla x_0 ustalonego punktu i h_1, h_2 dodatnich kroków. Dla $x \in S_h$ i naszego operatora różnicowego $L_h u_h(x) = -\sum_{k=1,2} \partial\overline{\partial}_k u_h(x) + c * u_h(x)$ definiujemy otoczenie siatkowe (ang. *mesh neighborhood*) punktu x jako podzbiór S_h taki, aby



Rysunek 7.3. Siatka dla obszaru nieprostokątnego. Czarne punkty należą do $\partial\Omega_h$.

$L_h u_h(x)$ było zdefiniowane poprzez wartości u_h w $N_h(x)$, czyli w tym przypadku otoczenie siatkowe $N_h(x)$ zawiera dany punkt x i cztery sąsiednie punkty leżące na pionowej i poziomej prostej przechodzącej przez x tzn. $N_h(x) = \{x, x + h_1 * \vec{e}_1, x - h_1 * \vec{e}_1, x + h_2 * \vec{e}_2, x - h_2 * \vec{e}_2, \}$, por. rysunek 7.4.

Wtedy definiujemy $\overline{\Omega}_h = S_h \cap \overline{\Omega}$, a

$$\Omega_h = \{x \in S_h \cap \Omega : N_h(x) \subset \overline{\Omega}\} \subset \Omega \cap S_h,$$

czyli złożoną z takich punktów Ω, że ten punkt i sąsiednie punkty na osiach siatki należą do Ω. Wtedy $\partial\Omega_h = \overline{\Omega}_h \setminus \Omega_h$, por. rysunek 7.3. (W przypadku bardziej skomplikowanych operatorów definicja $N_h(x)$ może być inna, a zatem i definicja Ω_h może być inna).

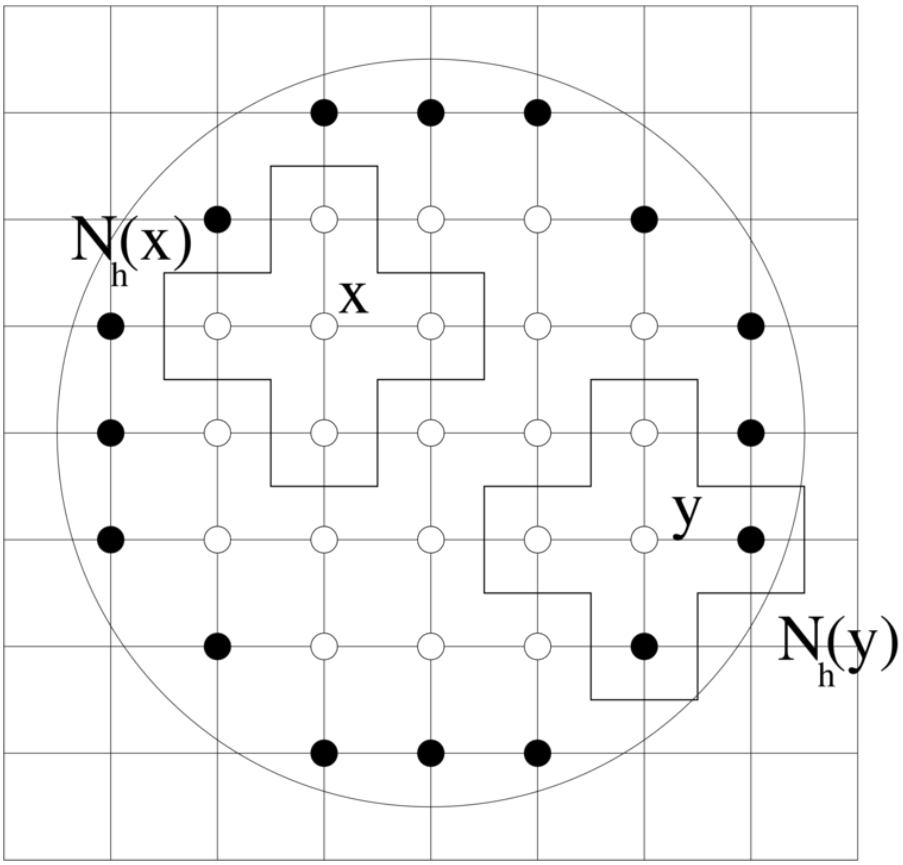
Następnie będziemy szukali funkcji zdefiniowanej na tej siatce tj. w $\overline{\Omega}_h = \Omega_h \cup \partial\Omega_h$.

Zauważmy, że wtedy dla każdego $x \in \Omega_h$

$$L_h u_h(x) = -\sum_{k=1,2} \partial\overline{\partial}_k u_h(x) + c u_h(x) = f_h(x) = f(x)$$

jest poprawnie zdefiniowany. Pojawia się natomiast pytanie, jak postawić warunek brzegowy Dirichleta w punktach $\partial\Omega_h$, które nie leżą na $\partial\Omega$. Najprostszym rozwiązaniem dla $x \in \partial\Omega_h$, który leży w Ω, jest postawienie w takim punkcie warunku:

$$l_h u_h(x) = g(x_s) = g_h(x),$$



Rysunek 7.4. Siatka dla obszaru nieprostokątnego. Otoczenia siatkowe dla operatora $L_h = -\sum_{k=1,2} \partial\overline{\partial}_k$.

gdzie $x_s \in \partial\Omega$ taki, że odległość od x jest nie większa niż h. Dalej postępujemy jak w przypadku obszaru prostokątnego.

Rozwiązanie wyjściowego zadania u jest określone we wszystkich punktach $\overline{\Omega}_h$, ponieważ $\partial\Omega_h \subset \overline{\Omega}$. Operator obcięcia definiujemy tak samo, tzn.:

$$r_h u(x) = u(x) \qquad x \in \overline{\Omega}_h.$$

Dla u dostatecznie gładkiego otrzymujemy:

$$\|L_h r_h u - f_h\|_{\infty,h,\Omega_h} := \max_{x\in\Omega_h} |L_h r_h u(x) - f_h(x)| = O(h^2)$$

dla $h = \max\{h_1, h_2\}$ i tylko:

$$\|l_h r_h u - g_h(x)\|_{\infty,h,\partial\Omega_h} := \max_{x\in\partial\Omega_h} |l_h r_h u(x) - g_h(x)| = O(h).$$

jeśli $\partial\Omega_h \not\subset \partial\Omega$. Pokazanie tego pozostawiamy jako zadanie.

Oznacza to, że dokładność schematu na rozwiązaniu wyjściowego zagadnienia różniczkowego, czyli rząd aproksymacji schematu (który formalnie zostanie zdefiniowany w kolejnym rozdziale, por. definicja 8.4) wynosi jeden. Istnieją oczywiście metody podwyższania rzędu aproksymacji w tym przypadku poprzez odpowiednią interpolację warunków brzegowych z brzegu obszaru na punkty siatki $\partial\Omega_h$ leżące wewnątrz Ω.

7.3. Zadania

Ćwiczenie 7.1. Udowodnij, (7.4).

Ćwiczenie 7.2. Rozpatrzmy zadanie różniczkowe jednowymiarowe:

$$-u''(x) + c(x)u(x) = f \qquad w \qquad [0,1]$$

czyli równanie (7.5), ale z warunkiem brzegowym Neumanna : tzn. z $u'(a) = g(a)$ i $u'(b) = g(b)$. Pokaż, że jeśli $c(x) \geqslant c_0 > 0$ to zadanie ma jednoznaczne rozwiązanie.

Rozważmy następującą dyskretyzację na siatce $\overline{\Omega}_h = \{x_k\}_{k=0,\ldots,N}$ z $h = (b-a)/N$ i $x_k = k*h$:

$$\begin{aligned} L_h u_h(x) &:= -\partial\overline{\partial} u_h(x) + c * u_h(x) = f(x) \qquad x \in \Omega_h \\ l_{h,a} u_h &:= \partial_h u(a) = g(a) \qquad l_{h,b} u_h = \overline{\partial}_h u(b) = g(b) \end{aligned}$$

Sprawdź, czy to zadanie dyskretne ma jednoznaczne rozwiązanie. Sformułuj je jako układ równań liniowych:

$$A\vec{u} = \vec{F}$$

dla $\vec{u} = (u_0, \ldots, u_N)^T$, znajdując macierz A i wektor prawej strony $\vec{F}$. Czy ta macierz jest symetryczna, czy jest nieosobliwa? Czy jest trój-diagonalna? Jak rozwiązać powyższy układ możliwie małym kosztem?

Ćwiczenie 7.3. Rozpatrzmy zadanie i schemat z poprzedniego zadania. Zbadaj, dla jakiego możliwie dużego p lokalny błąd schematu posiada rząd p, czyli

$$\max\{\max_{x\in\Omega_h} |L_h r_h u(x) - f(x)|, \max_{s\in\{a,b\}} |l_{h,s} r_h u(s) - g(s)|\} = O(h^p)$$

Zakładamy dowolnie wysoką regularność rozwiązania zadania wyjściowego.

Ćwiczenie 7.4. Rozpatrzmy zadanie różniczkowe i schemat z ćwiczenia 7.2, ale z $c = \alpha = \beta = 0$. Pokaż, że zadanie nie ma jednoznacznego rozwiązania w ogólności, ale ma z dokładnością do stałej, o ile $\int_a^b f = 0$. Jaki warunek musi spełniać f_h, aby zadanie dyskretne miało rozwiązanie? Sformułuj ten schemat jako układ równań liniowych, jak w ćwiczeniu 7.2. Pokaż, że jądro macierzy A jest jednowymiarowe. Znajdź bazę jądra tej macierzy. Wykorzystując tę informację zaproponuj tanią (w sensie ilości operacji arytmetycznych) metodę znalezienia rozwiązania dyskretnego z dodatkowym warunkiem $u_h(a) = 0$.

Ćwiczenie 7.5. Analogicznie do przypadku jednowymiarowego, biorąc $x_{kl} = (k*h, l*h)$,

$$u_{k-1+(l-1)*(N-1)} = u(x_{kl})$$

i $\vec{u} = (u_{k-1+(l-1)*(N-1)})_{k,l=1}^{N-1}$ możemy zapisać zadanie dyskretne (7.11), jako układ równań liniowych $A_h\vec{u} = \vec{f}$ z macierzą A_h i wektorem prawej strony $\vec{f}$. Wyznacz tę macierz i ten wektor (por. (7.7) dla przypadku jednowymiarowego). Oblicz, ile elementów różnych od zera ma ta macierz. Policz, ile operacji arytmetycznych jest potrzebnych do rozwiązania tego układu równań liniowych z zastosowaniem metody Choleskiego, czyli rozkładu $A = L^T L$ dla L macierzy dolnotrójkątnej w wersji dla macierzy pasmowych.

Ćwiczenie 7.6 (laboratoryjne)**.** Stwórz w octavie macierz z poprzedniego zadania dla $c = 0$, tzn. macierz układu równań powstałego z (7.11), jako macierz pełną i rzadką (można wykorzystać funkcję octave'a **sparse**()). Następnie rozwiąż ten układ dla wektora prawej strony odpowiadającego funkcji $f = -\triangle(u_*)$ dla $u_* = x*(1-x)*y(1-y)$.

1. Korzystając z narzędzi octave'a **tic**() i **toc**() sprawdź czas rozwiązywania dla różnych N np. $N = 20, 80, 320$ itp. dla obu typów macierzy. Czy różnica jest znacząca?
2. Zbadaj błąd dyskretny w normie maksimum (funkcja octave'a **norm**(wektor,'inf')) i w dyskretnej normie L^2 (np. używając funkcji octave'a **norm**(x,2)) między rozwiązaniem dokładnym u_*, a rozwiązaniem dyskretnym dla $N = 10, 20, 40, 80, 160$. Czy eksperyment potwierdza teorię, tzn. czy dla podwojonych N (czyli połowionych h) błąd maleje czterokrotnie?

Ćwiczenie 7.7 (laboratoryjne)**.** Rozwiąż równanie $-\frac{d^2u}{dx^2} + u = 0$ z warunkami brzegowymi $u(0) = u(20) = 1$ na $[0, 20]$ przy pomocy metody różnic skończonych, tzn. rozwiąż zadanie (7.1) dla $a = 0, b = 20$ i $c(x) = 1$ za pomocą schematu (7.5) dla $N = 100, 200$. Policz normę maksimum w punktach siatki między dokładnymi wartościami rozwiązania $u^*(x) = (e^{t-20}+e^{-t})/(1+e^{-20})$ a rozwiązaniem dyskretnym u_h zadania (7.5), Porównaj z wynikami metody strzałów zastosowanej do tego zadania tj. z ćwiczeniem 6.2.

Ćwiczenie 7.8 (częściowo laboratoryjne)**.** Rozpatrzmy przeskalowaną macierz z (7.7) (dla $c = 0$):

$$A_{N-1} = \begin{pmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{pmatrix} \in \mathbb{R}^{N-1,N-1}$$

Pokaż, że jej wartości własne to $\lambda_k = 4 * \sin^2(\frac{k*\pi}{2*N})$ dla $k = 1, \ldots, N-1$ z odpowiednimi wektorami własnymi: $\vec{x}_k = \{\sin(\frac{k*\pi*j}{N})\}_{n=1}^{N-1}$ dla $k = 1, \ldots, N-1$. Oszacuj uwarunkowanie macierzy tzn. $\frac{\max_k \lambda_k}{\min_k \lambda_k}$ dla $N >> 1$.

Porównaj z wynikami otrzymanymi przy pomocy funkcji octave'a **eig**() i **cond**().

Wskazówka. Korzystamy z wzorów trygonometrycznych: $\sin(a) + \sin(b) = 2 * \sin((a+b)/2) * \cos((a-b)/2)$ *oraz* $\cos(2*a) = 1-2*\sin^2(a)$. *Rozwiązanie.* Zauważmy, że biorąc $v_{j,k} = (v_k)_j = \sin(\frac{k*\pi*j}{N})$ otrzymujemy, że $v_{0,k} = v_{N,k} = 0$. Zatem wystarczy sprawdzić, czy zachodzą równania $-v_{j-1,k} + 2v_{j,k} - v_{j+1,k} = \lambda_k v_{j,k}$ dla $j = 1, \ldots, N-1$.
Biorąc $\alpha_k = k * \pi/N$ otrzymujemy:

$$\begin{aligned} -v_{j-1,k} + 2v_{j,k} - v_{j+1,k} &= -(\sin((j-1)\alpha_k) - \sin((j+1)\alpha_k) + 2 * \sin(j * \alpha_k) \\ &= -2 * \sin(j * \alpha_k) * \cos(\alpha_k) + 2 * \sin(j * \alpha_k) \\ &= 2 * (1 - \cos(\alpha_k)) \sin(j * \alpha_k) = 4 * \sin^2(\alpha_k/2) \sin(j * \alpha_k) \\ &= 4 * \sin^2(\alpha_k/2) * v_{j,k} = 4 * \sin^2(\frac{k*\pi}{2*N}) * v_{j,k} = \lambda_k * v_{j,k}. \end{aligned}$$

Uwarunkowanie $\frac{\sin^2((N-1)*\pi)/2*N)}{\sin^2(\pi/2*N)}$ dla dużych N dąży do nieskończoności jak N^2.

8. Teoria zbieżności schematów różnicowych

W tym rozdziale przedstawimy ogólną teorię zbieżności schematów różnicowych, a następnie pokażemy m.in. zastosowanie tej teorii do przykładów z poprzedniego wykładu.

Osoby zainteresowane obszerniejszym przedstawieniem teorii różnic dzielonych odsyłamy do monografii [27].

8.1. Ogólna teoria zbieżności schematów różnicowych

W tym podrozdziale opiszemy ogólną teorię zbieżności schematów różnicowych. Ograniczymy się do szczegółowego omówienia przypadku schematów liniowych, tzn. aproksymacji równań różniczkowych liniowych.

Teoria ta potrzebna jest zarówno do badania zbieżności schematów różnicowych dla równań eliptycznych, jak i dla schematów dla innych typów równań, np. równań parabolicznych.

Załóżmy, że rozpatrujemy następujące zadanie różniczkowe: chcemy znaleźć u funkcję określoną na obszarze $\overline{\Omega}$ taką, że spełnia równanie różniczkowe z warunkami brzegowymi:

$$\begin{aligned} Lu(x) &= f(x) \qquad x \in \Omega && (8.1)\\ l_k u(x) &= g_k(x) \qquad x \in \Gamma_k, \qquad k = 1, \ldots, s, && (8.2) \end{aligned}$$

gdzie f, g_k - to dane funkcje, L - to operator różniczkowy liniowy, l_k - to odpowiedni operator różniczkowy brzegowy liniowy określony na Γ_k dla $\Gamma_k \subset \partial\Omega$.

Będziemy zakładać, że powyższe zadanie jest poprawnie postawione, tzn. że ma jednoznaczne rozwiązanie $u \in U$ dla U przestrzeni liniowej funkcji określonych na $\overline{\Omega}$ z normą $\|\cdot\|_U$. Zakładamy też, że

$$L : U \to F$$

dla F przestrzeni funkcji określonych na Ω, a

$$l_k : U \to \Phi_k \qquad k = 1, \ldots, s$$

dla Φ_k przestrzeni funkcji określonych na Γ_k. Wyjściowe zadanie różniczkowe możemy zapisać w postaci operatorowej jako: znaleźć $u \in U$ takie, że

$$\begin{aligned} Lu &= f && (8.3)\\ l_k u &= g_k \qquad k = 1, \ldots, s. && (8.4) \end{aligned}$$

Zdefiniujmy $\overline{\Omega}_h$ jako siatkę, tzn. zbiór punktów izolowanych, węzłów należących do $\overline{\Omega}$ z parametrem h.

Zakładamy, że istnieje rodzina siatek $\{\overline{\Omega}_h\}_h$, czyli rodzina zbiorów punktów izolowanych należących do $\overline{\Omega}$ indeksowanych parametrem h, należącym do pewnego zbioru $\omega \subset (0, h_0] \subset \mathbb{R}_+$ takim, że $0 \in \overline{\omega}$ (tzn. że istnieje podciąg siatek Ω_{h_k} taki, że $h_k \to 0$).

W praktyce najczęściej stosuje się siatki równomierne, tzn. podzbiory $a + h*\mathbb{Z}^d$ dla ustalonego punktu $a \in \mathbb{R}^d$. Ewentualnie stosuje się siatki o jednolitych krokach w danym kierunku w $\mathbb{R}^d$.

Siatkę $\overline{\Omega}_h$ przedstawiamy w postaci $\overline{\Omega}_h = \Omega_h \cup \partial\Omega_h$, gdzie Ω_h będziemy nazywać zbiorem punktów siatkowych wewnętrznych (zazwyczaj zawartych w Ω), a $\partial\Omega_h$ - zbiorem punktów siatkowych brzegowych (zawartych albo leżących w pobliżu $\partial\Omega$). W zbiorze $\partial\Omega_h$ punktów brzegowych możemy dalej wyróżniać podzbiory $\Gamma_{k,h}$. Zakładamy, że rodzina siatek $\{\overline{\Omega}_h\}_h$ jest gęsta w sensie następującej definicji:

Definicja 8.1. Rodzina siatek $\{\overline{\Omega}_h\}_h$ jest **gęsta** (ang. *dense*) w $\overline{\Omega}$, gdy dla dowolnego $\epsilon > 0$ istnieje $h_1 \in \omega$ takie, że dla $h < h_1$ i dowolnego $x \in \overline{\Omega}$ kula $K(x, \epsilon)$ o środku w x i promieniu ϵ zawiera co najmniej jeden punkt $y \in \overline{\Omega}_h$.

Proszę zauważyć, że rodzina siatek zdefiniowana w Rozdziale 7.1 jest w sposób oczywisty gęsta dla $[a, b]$.

Wprowadzamy teraz rodzinę zadań przybliżonych (schematów różnicowych), które dają się zapisać w następujący sposób: chcemy znaleźć funkcję u_h określoną na $\overline{\Omega}_h$ taką, że

$$\begin{aligned} L_h u_h(x) &= f_h(x) \qquad x \in \Omega_h \\ l_{k,h} u(x) &= g_{k,h}(x) \qquad x \in \Gamma_{k,h}, \qquad k = 1, \ldots, s, \end{aligned}$$

czy inaczej - operatorowo

$$L_h u_h = f_h \tag{8.5}$$
$$l_{k,h} u = g_{k,h} \qquad k = 1, \ldots, s. \tag{8.6}$$

Zakładamy, że $L_h : U_h \to F_h$ i $l_{k,h} : U_h \to \Phi_{k,h}$ dla $k = 1, \ldots, s$, gdzie:
1. U_h jest przestrzenią liniową unormowaną funkcji określonych na $\overline{\Omega}_h$ z normą $\|\cdot\|_{U_h}$,
2. F_h jest przestrzenią liniową unormowaną funkcji określonych na Ω_h z normą $\|\cdot\|_{F_h}$,
3. $\Phi_{k,h}$ jest przestrzenią liniową unormowaną funkcji określonych na $\Gamma_{k,h}$ z normą $\|\cdot\|_{\Phi_{k,h}}$.

Zazwyczaj wszystkie rozpatrywane przestrzenie są zupełne, tzn. są przestrzeniami Banacha. W przypadku gdy Ω jest ograniczony, są one też przestrzeniami skończenie wymiarowymi. Jeśli L_h i wszystkie $l_{k,h}$ są operatorami liniowymi, to mówimy, że rozpatrujemy zadanie przybliżone (dyskretne) liniowe, czy schemat różnicowy liniowy. W przeciwnym razie - gdy choć jeden z operatorów jest nieliniowy, to mamy do czynienia z zadaniem przybliżonym nieliniowym, czy schematem różnicowym nieliniowym.

Proszę zauważyć, że rozpatrujemy rodzinę zadań przybliżonych, parametryzowanych przez h. Tak, jak w przykładzie w Rozdziale 7.1, aby mówić o zbieżności rozwiązania zadania dyskretnego $u_h \in U_h$ do $u \in U$ musimy mieć możliwość porównania obu funkcji. Dlatego zakładamy, że istnieje rodzina operatorów obcięcia (ang. *restriction*) $r_h^U : U \to U_h$, które są liniowe i ograniczone jednostajnie (ang. *uniformly bounded*) względem h, tzn. $\exists K > 0 \quad \forall h \in \omega \quad \forall u \in U$

$$\|r_h^U u\|_{U_h} \leqslant K \|u\|_U.$$

Operator obcięcia pozwala porównywać rozwiązania w normie przestrzeni dyskretnej, ale możemy porównywać je również w normie przestrzeni U. W tym celu musimy wprowadzić rodzinę operatorów liniowych przedłużenia $p_h^U : U_h \to U$. Najczęściej za operatory przedłużenia bierze się odpowiednie operatory interpolacji.

Uwaga 8.1. Można wprowadzić pojęcie zbieżności aproksymacji przestrzeni. Tzn. rodzinę trójek (u_h, r_h^U, p_h^U) nazywamy aproksymacją przestrzeni U i mówimy, że ta aproksymacja jest zbieżna, jeśli dla dowolnego $u \in U$ zachodzi zbieżność

$$\|p_h^U r_h^U u - u\|_U \to 0 \qquad h \to 0.$$

W teorii zbieżności metod różnicowych najczęściej nie stosuje się operatorów przedłużenia, a za to wprowadza się warunek zgodności norm:

Definicja 8.2. Jeżeli dla danej przestrzeni unormowanej $(U, \|\cdot\|_U)$ i rodziny przestrzeni unormowanych z odpowiednimi operatorami obcięcia $(U_h, \|\cdot\|_{U_h}, r_h^U)$ zachodzi zbieżność

$$\lim_{h\to 0} \|r_h^U u\|_{U_h} = \|u\|_U \qquad \forall u \in U,$$

to mówimy, że normy dyskretne $\|\cdot\|_{U_h}$ są zgodne (ang. *consistent*) z normą $\|\cdot\|_U$,

Od tej pory będziemy zakładali zgodność norm dyskretnych z normą w U, według powyższej definicji.

Definicja 8.3. Zadanie przybliżone (zadanie dyskretne, schemat różnicowy) (8.5)-(8.6) jest zbieżne (czasami używa się terminu zbieżne dyskretnie) jeśli

$$\|r_h^U u - u_h\|_{U_h} \to 0 \qquad h \to 0,$$

gdzie u - to rozwiązanie zadania (8.3)-(8.4), a $u_h \in U_h$ - to rozwiązanie dyskretne zadania przybliżonego (8.5)-(8.6).

Jeśli dodatkowo zachodzi

$$\|r_h^U u - u_h\|_{U_h} \leqslant O(h^p)$$

to mówimy o zbieżności (dyskretnej) rzędu p.

Wielkość $\|r_h^U u - u_h\|_{U_h}$ będziemy nazywać błędem dyskretnym dla zadania przybliżonego (ang. *dicrete error*).

Kolejnym krokiem jest wprowadzenie pojęcia aproksymacji zadania ciągłego (wyjściowego zadania różniczkowego) przez zadanie dyskretne.

Definicja 8.4 (aproksymacja; rząd schematu; (ang. *consistency*))**.** Mówimy, że zadanie przybliżone (8.5)-(8.6) aproksymuje zadanie (8.3)-(8.4), jeśli lokalne błędy aproksymacji zdefiniowane jako

$$e_{0,h} := \|L_h r_h^U u - f_h\|_{F_h} \qquad e_{k,h} = \|l_{k,h} r_h^U u - g_{k,h}\|_{\Phi_{k,h}} \qquad k = 1, \ldots, s,$$

dążą do zera dla $h \to 0$. Tutaj u jest rozwiązaniem zadania (8.3)-(8.4), a $f_h, g_{k,h}$ są z zadania dyskretnego (8.5)-(8.6). Jeśli dodatkowo zachodzi:

$$e_{k,h} = O(h^p) \qquad k = 0, 1, \ldots, s,$$

to mówimy, że schemat aproksymuje (8.3)-(8.4) z rzędem p (ang. *local truncation error is of order p*), (inaczej, że lokalne błędy aproksymacji są rzędu p, dane zadanie przybliżone lub schemat różnicowy ma rząd p, rząd aproksymacji schematu wynosi p).

Drugim ważnym pojęciem jest stabilność zadania dyskretnego. Tu podamy definicje stabilności dla schematu liniowego:

Definicja 8.5 (stabilność; (ang. *stability*))**.** Liniowe zadanie przybliżone (8.5)-(8.6) jest stabilne (poprawnie postawione), jeśli istnieje stała h_0 taka, że dla dowolnego $h \in \omega$, $h \leqslant h_0$ i dla dowolnych $f_h \in F_h$ i $g_{k,h} \in \Phi_{k,h}$ $k = 1, \ldots, s$ zachodzą:

1. istnieje jednoznacznie wyznaczone rozwiązanie $u_h \in U_h$ spełniające (8.5)-(8.6),
2. rozwiązanie to spełnia następującą nierówność:

$$\|u_h\|_{U_h} \leqslant C(\|f_h\|_{F_h} + \sum_{k=1}^{S} \|g_{k,h}\|_{\Phi_{k,h}},$$

gdzie C - to dodatnia stała niezależna od h (ani oczywiście od $f_h, g_{k,h}$).

Uwaga 8.2. W literaturze czasami za stabilność zadania przybliżonego przyjmuje się tylko warunek (2) z definicji 8.5.

Proszę zauważyć, że stabilność zadania przybliżonego jest samoistną cechą związaną tylko z definicją samego zadania dyskretnego- ona nie zależy w żaden sposób od rozwiązania równania różniczkowego. Dodatkowo warto też zauważyć, że jeśli U_h jest przestrzenią skończenie wymiarową, istnieje rozwiązanie (8.5)-(8.6) i spełniony jest warunek (2) z definicji 8.5, to wtedy to rozwiązanie jest jednoznaczne.

Sformułujemy teraz następujące twierdzenie o zbieżności zadania przybliżonego:

Twierdzenie 8.1 (Lax-Filipow)**.** *Jeśli liniowe zadanie przybliżone (8.5)-(8.6) jest stabilne oraz aproksymuje zadanie (8.3)-(8.4), którego rozwiązaniem jest u, wtedy zadanie przybliżone jest zbieżne i*

$$\|r_h^U u - u_h\|_{U_h} \leqslant C(\sum_{k=0}^{s} e_{k,h}).$$

Tutaj u_h - to rozwiązanie zadania przybliżonego (8.5)-(8.6).

Z powyższego twierdzenia otrzymujemy od razu następujący wniosek:

Wniosek 8.1. *Jeśli zadanie przybliżone (8.5)-(8.6) jest stabilne oraz aproksymuje zadanie (8.3)-(8.4) z rzędem p, to*

$$\|r_h^U u - u_h\|_{U_h} = O(h^p).$$

Dowód. Oznaczmy $z_h = r_h^U u - u_h$. Z liniowości L_h i $l_{k,h}$ dla $k = 1, \ldots, s$ wynika, że z_h spełnia zadanie przybliżone z odpowiednimi prawymi stronami:

$$L_h z_h = L_h r_h^U - f_h = w_h, \qquad l_{k,h} z_h = l_{k,h} r_h^U u - g_{k,h} = w_{k,h} \qquad k = 1, \ldots, s,$$

zatem z definicji stabilności zadania przybliżonego otrzymujemy następujące oszacowanie:

$$\|z_h\|_{U_h} \leqslant C(\|g_h\|_{F_h} + \sum_{k=1}^{s} \|w_{k,h}\|_{\Phi_{k,h}}) = C(\|L_h r_h^U - f_h\|_{F_h} + \sum_{k=1}^{s} \|l_{k,h} r_h^U u - g_{k,h}\|_{\Phi_{k,h}}).$$

Następnie z faktu aproksymacji zadania (8.3)-(8.4) przez zadanie przybliżone (8.5)-(8.6) otrzymujemy ostatecznie oszacowanie:

$$\|r_h^U u - u_h\|_{U_h} = \|z_h\|_{U_h} \leqslant C(\sum_{k=0}^{s} e_{k,h}) = O(h^p) \to 0 \qquad h \to 0.$$

□

Powyższe twierdzenie można krótko podsumować, że aby otrzymać schemat zbieżny z rzędem p musi być on stabilny i posiadać rząd aproksymacji p.

Proszę zauważyć, że twierdzenie jest bardzo ogólne, a dowód jest prosty. Pojawia się pytanie: jak dobrać odpowiednie przestrzenie i operatory, aby zadania przybliżone (schematy) były stabilne i miały możliwie wysoki rząd aproksymacji.

Uwaga 8.3. Proszę zauważyć, że powyższa teoria zbieżności może zostać zastosowana do zadań różniczkowych rożnego typu - zarówno eliptycznych, jak i parabolicznych, czy hiperbolicznych.

8.2. Zastosowanie teorii zbieżności do prostych schematów jedno- i dwuwymiarowych

8.2.1. Przypadek jednowymiarowy

Wracamy teraz do dyskretyzacji modelowego zadania jednowymiarowego (7.5)-(7.6). Za przestrzeń U weźmy przestrzeń funkcji ciągłych na $\overline{\Omega}$, czyli $C([a,b])$ z normą supremum $\|u\|_\infty = \sup_{t\in[a,b]} |u(t)|$. Oznaczmy przez $C_h(K_h)$ zbiór funkcji określonych na dowolnym podzbiorze K_h siatki $\overline{\Omega}_h$ z normą $\|u\|_{\infty,h,K_h} = \max_{x\in K_h} |u(x)|$. Za przestrzeń dyskretną przyjmijmy $U_h = C(\overline{\Omega}_h)$. Jeżeli wprowadzimy operatory $L_h : U_h \to F_h$ i $l_h : U_h \to \Phi_h$ zdefiniowane jako (por. (7.3)):

$$\begin{aligned} L_h u_h(x) &= -\partial\overline{\partial} u_h(x) + c * u_h(x) \qquad x \in \Omega_h \\ l_h u_h(x) &= u(x) \qquad x \in \partial\Omega_h = \overline{\Omega}_h \setminus \Omega_h \end{aligned}$$

dla $F_h = C_h(\Omega_h)$ i $\Phi_h = C_h(\partial\Omega_h)$, to zadanie (7.5)-(7.6) możemy zapisać w formie operatorowej jako:

$$\begin{aligned} L_h u_h &= f_h, \\ l_h u_h &= g_h \end{aligned} \tag{8.7}$$

dla $f_h \in F_h$ zdefiniowanego jako $(f_h)(x) = f(x)$ dla $x \in \Omega_h$ oraz $g_h \in \Phi_h$ z $(g_h)(x) = g(x)$ dla $x \in \partial\Omega_h$.

W tym przypadku dla funkcji ciągłej przekształceniem obcięcia (ang. *restriction*) jest $r_h^{C([a,b])} : U \to U_h$ zdefiniowany jako

$$(r_h^{C([a,b])} u)(x) = r_h u(x) = u(x) \qquad x \in \overline{\Omega}_h.$$

Możemy teraz zbadać zbieżność błędu dyskretnego:

$$\|r_h u - u_h\|_{\infty,h}$$

dla $h \to 0$, co jest równoważne badaniu zbieżności w punktach siatki. Zauważmy, że otrzymujemy

$$\|r_h u\|_{\infty,h} \leqslant \|u\|_\infty \qquad \forall u \in U,$$

co oznacza jednostajną ograniczoności operatorów obcięcia.

Można też w tym przypadku łatwo wprowadzić operator przedłużenia (ang. prolongation) $p_h : U_h \to U$. Np. niech p_h będzie funkcją ciągłą liniowo interpolującą wartości u_h pomiędzy punktami siatki tj.

$$p_h(x) = u_h(x_k) + \frac{u_h(x_{k+1}) - u_h(x_k)}{h}(x - x_k) \qquad x_k \leqslant x \leqslant x_{k+1}.$$

Następnie możemy badać zbieżność błędu $\|p_h u_h - u\|_\infty$ dla $h \to 0$. Jeśli błąd zbiega do zera, to mówimy o zbieżności schematu w normie supremum.

Zauważmy, że w naszym przypadku dodatkowo zachodzi

$$\|p_h r_h u - u\|_\infty \to 0 \qquad h \to 0 \qquad \forall u \in U, \tag{8.8}$$

czyli zbieżność aproksymacji przestrzeni wyjściowej przez przestrzeń dyskretną oraz

$$\lim_{h\to 0} \|r_h u\|_{\infty,h} = \|u\|_\infty \qquad \forall u \in U, \tag{8.9}$$

czyli zachodzi zgodność rodziny norm przestrzeni dyskretnych $(U_h, \|\cdot\|_{\infty,h})$ z normą przestrzeni wyjściowej $(U, \|\cdot\|_\infty)$. Wykazanie tego pozostawiamy jako zadanie, por. ćwiczenie 8.1.

Innym wyborem przestrzeni i norm jest badanie zbieżności i błędu w normie $L^2(a,b)$, czy odpowiednio dyskretnej normie L^2_h definiowanej dla $u \in U_h$ jako:

$$\|u\|_{0,h} = \sqrt{h \sum_{x \in \overline{\Omega}_h} |u(x)|^2},$$

gdzie $U_h = L^2_h(\overline{\Omega}_h)$ - to przestrzeń wszystkich funkcji określonych na $\overline{\Omega}_h$. Oczywiście zmieniliśmy oznaczenie przestrzeni funkcji dyskretnych na siatce. Jest to ten sam zbiór funkcji określonych na siatce, ale zmieniła się norma dyskretna.

Aby otrzymać zgodność norm powinniśmy inaczej zdefiniować obcięcia np. poprzez uśrednienia, czyli $r_h^{L^2} : L^2(\Omega) \to L^2_h(\overline{\Omega}_h)$ dla $x \in \overline{\Omega}_h$ definiujemy:

$$r_h^{L^2} u(x) = \frac{1}{K(x,h) \cap \Omega} \int_{B(x,h)\cap\Omega} u(x)\, dx.$$

dla $K(x,h)$ kuli o środku w x i promieniu h.

Inna możliwość to rozważenie zbioru funkcji ciągłych $U = C([a,b])$ ale z normą L^2, oraz normy dyskretnej typu L^2_h na U_h. Następnie możemy przeprowadzić analizę z obcięciem $r_h := r_h^{C([a,b])} u$. Zbiór funkcji $U = C([a,b])$ z normą L^2 nie jest przestrzenią zupełną, ale jest gęstą podprzestrzenią przestrzeni $L^2(a,b)$.

Nietrudno zauważyć, że problem przybliżony aproksymuje problem wyjściowy z rzędem dwa, o ile rozwiązanie należy do $C^4(\overline{\Omega})$, w obu powyżej przedstawionych przestrzeniach dyskretnych, czyli w odpowiednich normach dyskretnych. Wykazanie, że schemat jest stabilny zarówno w $C(\overline{\Omega}_h)$ jak i $L^2(\overline{\Omega}_h)$ jest trudniejsze. Zajmiemy się tym w kolejnych wykładach.

8.2.2. Przypadek dwuwymiarowy

Rozpatrzmy ponownie modelowe zadanie dwuwymiarowe na kwadracie jednostkowym (7.8). Analogicznie, jak w przypadku jednowymiarowym, niech $U = C([0,1]^2)$ z normą supremum $\|u\|_\infty = \sup_{t\in[0,1]^2} |u(t)|$ i $C_h(K_h)$ będzie przestrzenią funkcji określonych na podzbiorze K_h siatki $\overline{\Omega}_h$ (por. (7.10)) z normą $\|u\|_{\infty,h,K_h} = \max_{x\in K_h} |u(x)|$. Przestrzeń dyskretną definiujemy jako $U_h = C(\overline{\Omega}_h)$.

Operator siatkowy (ang. *mesh operator or discrete operator*) $L_h : U_h \to F_h$ i brzegowy $l_h : U_h \to \Phi_h$ definiujemy jako:

$$\begin{aligned} (L_h u_h)(x) &= (-\sum_{k=1,2} \partial\overline{\partial}_k + c) u_h(x), \qquad x \in \Omega_h \\ (l_h u_h)(x) &= u(x), \qquad x \in \partial\Omega_h = \overline{\Omega}_h \setminus \Omega_h \end{aligned}$$

dla $F_h = C_h(\Omega_h)$ i $\Phi_h = C_h(\partial\Omega_h)$. Teraz zadanie (7.11) możemy zapisać w formie operatorowej jako

$$\left\{ \begin{array}{rcl} L_h u_h &=& f_h \\ l_h u_h &=& g_h \end{array} \right. \tag{8.10}$$

dla funkcji prawej strony $f_h \in F_h$ oraz $g_h \in \Phi_h$ zdefiniowanych jako $f_h(x) = f(x)$ dla $x \in \Omega_h$ i $g_h(x) = g(x)$ dla $x \in \partial\Omega_h$. Operatorem obcięcia (ang. *restriction*) jest $r_h : U \to U_h$, zdefiniowany jako

$$r_h u(x) = u(x) \qquad x \in \overline{\Omega}_h.$$

Tak samo jak w przypadku jednowymiarowym badamy błąd: $\|r_h u - u_h\|_{\infty,h}$, lub w dyskretnej normie L^2, tzn. w

$$\|u\|_{0,h} = \sqrt{h^2 \sum_{x \in \overline{\Omega}_h} |u(x)|^2}.$$

Tu $U_h = L_h^2(\overline{\Omega}_h)$ jest zdefiniowana jako przestrzeń wszystkich funkcji określonych na $\overline{\Omega}_h$. Oczywiście zbiór funkcji siatkowych jest ten sam, zmieniła się tylko norma.

Można pokazać, że zachodzi zgodność norm dyskretnych z odpowiednimi normami, oraz że schemat (7.11) posiada rząd aproksymacji dwa i jest stabilny w obu normach dyskretnych. Wykazanie rzędu aproksymacji jest prostym zadaniem, natomiast pokazanie stabilności jest trudniejsze, por. rozdziały 9 i rozdział 10.

8.3. Zadania

Ćwiczenie 8.1. Udowodnij (8.8) i (8.9).

Ćwiczenie 8.2. Zbadaj rząd lokalnych błędów aproksymacji schematu (8.7) dyskretyzacji modelowego problemu jednowymiarowego w obu normach dyskretnych.

Ćwiczenie 8.3. Wykaż, że rząd aproksymacji schematu (7.11) w dyskretnych normach maksimum i L_h^2 wynosi dwa, o ile rozwiązania wyjściowego zadania różniczkowego są dostatecznie gładkie.

Ćwiczenie 8.4. Zbadaj rząd lokalnych błędów aproksymacji schematu (8.10) dyskretyzacji modelowego problemu dwuwymiarowego w obu normach dyskretnych.

Ćwiczenie 8.5. (Przybliżony warunek brzegowy) Rozpatrzmy modelowe zadanie jednowymiarowe z warunkiem brzegowym Dirichleta:

$$-u''(x) = f(x) \qquad x \in (0,1) \qquad x(0) = a \quad x(1) = b$$

dla $f \in C^\infty$.

Rozpatrzmy następującą dyskretyzację zbudowaną na siatce $\overline{\Omega}_h = \{x_k\}_{k=0,\ldots,N-1}$ dla $x_k = k * h$ z $k = 0, \ldots, N-1$ z $(N-1) * h < 1 < N * h$. Definiujemy $\Omega_k = \{k * h\}_{k=1,\ldots,N-2}$ i $\partial\Omega_h = \{0, (N-1) * h\}$, oraz operatory:

$$L_h u_h(x) = \partial\overline{\partial} u_h(x) \qquad x \in \Omega_h$$

i $l_h u_h(0) = a$, $l_h u_h(x_{N-1}) = b$. Zbadaj rząd lokalnego błędu aproksymacji tej dyskretyzacji w dyskretnej normie maksimum.

Wskazówka. Wystarczy zbadać błąd operatora brzegowego w punkcie x_{N-1}. W pozostałych punktach błąd jest jak w schemacie (8.7).

Ćwiczenie 8.6. Rozważmy modelowe zadanie, jak i siatkę niezawierającą prawy koniec obszaru, tak jak w poprzednim ćwiczeniu.

Operatory L_h i $l_h(0) = a$ definiujemy tak samo, natomiast zmodyfikujmy operator brzegowy l_h w prawym końcu, tzn. w punkcie $x_r := x_{N-1} < 1$.

Rozpatrzmy tzw. aproksymację Collatza, tzn. niech wartość $l_h(x_r) = b$ będzie liniowo interpolowała warunek brzegowy w końcu obszaru:

$$l_h u_h(x_r) = u_h(x_r) + \frac{u_h(x_r) - u_h(x_r - h)}{h}(1 - x_r) = (1 + \frac{\tilde{h}}{h})u_h(x_r) - \frac{\tilde{h}}{h}u_h(x_r - h) = b,$$

gdzie $\tilde{h} = 1 - x_r < h$.

Zbadaj lokalny błąd aproksymacji tego schematu w normach dyskretnych maksimum i L_h^2 i jego rząd, tzn. czy zachowuje się jak $O(h^p)$ dla pewnego p naturalnego.

Ćwiczenie 8.7. Rozpatrzmy zadanie z poprzedniego ćwiczenia, ale z siatką nierównomierną: $\overline{\Omega}_h = \{x_k\}_{k=0,\ldots,N-1} \cup \{1\}$ dla $x_k = k * h$ z $k = 0, \ldots, N-1$ z $(N-1) * h < 1 < N * h$. Operator l_h możemy zdefiniować jako

$$l_h u(0) = l_h u(\tilde{x}) = 0,$$

gdzie $\tilde{x} = 1$, ale musimy zmodyfikować definicję L_h, tzn.

$$L_h u(x_k) = -\partial\overline{\partial} u(x_k) \qquad k = 1, \ldots, N-2$$

i

$$L_h u(x_{N-1}) = a u(x_{N-2}) + b u(x_{N-1}) + c u(\tilde{x}).$$

Wyznacz a, b, c w zależności od wartości h i $\tilde{h} = 1 - x_{N-1} < h$ tak, aby lokalny błąd aproksymacji schematu był możliwie mały.

Ćwiczenie 8.8. Rozpatrzmy zadanie jednowymiarowe $-\frac{d^2u}{dt^2} + u = f$ na $[(0,1)$ z warunkiem Neumanna $\frac{du}{dt}(0) = \frac{du}{dt}(1) = 0$. Rozpatrzmy następującą dyskretyzację zbudowaną na siatce $\overline{\Omega}_h = \{x_k\}_{k=0,\ldots,N-1}$ dla $x_k = k * h$ z $h = 1/N$.

$$L_h u(x_k) = -\partial\overline{\partial}_h u(x_k) + u(x_k) = f(x_k) \qquad k = 1, \ldots, N-1$$

oraz

$$l_1 u(0) = \partial_h u(0) = 0, \qquad l_2 u(0) = \overline{\partial}_h u(1) = 0.$$

Zbadaj rząd lokalnego błędu aproksymacji tego schematu względem parametru siatki h w dyskretnej normie maksimum i dyskretnej normie L_h^2.

Ćwiczenie 8.9. Rozpatrzmy zadanie jednowymiarowe $-\frac{d^2u^*}{dt^2} + u^* = f$ na $(0,1)$ z warunkiem Neumanna $\frac{du^*}{dt}(0) = \frac{du^*}{dt}(1) = 0$. Rozpatrzmy następującą dyskretyzację o podwyższonym rzędzie zbudowaną na siatce $\overline{\Omega}_h = \{x_k\}_{k=0,\ldots,N-1}$ dla $x_k = k * h$ z $h = 1/N$. W punktach wewnętrznych siatki stosujemy standardowo aproksymacje na trzech punktach:

$$L_h u(x_k) = -\partial\overline{\partial}_h u(x_k) + u(x_k) = f(x_k) \qquad k = 1, \ldots, N-1$$

natomiast na brzegu podnosimy rząd schematu, a dokładniej zakładamy, że równanie jest spełnione w punktach brzegu, tzn. funkcja f jest określona na $\overline{\Omega} = [0,1]$ i $-\frac{d^2u^*(x)}{dt^2} + u^*(x) = f(x)$ dla $x \in \{0,1\}$. Rozpatrzmy lewy punkt brzegu $x = 0$. Widzimy, że

$$\partial_h u^*(0) = \frac{du^*}{dt}(0) + \frac{d^2u}{dt^2}(0) * \frac{h}{2} + O(h^2) = \frac{du^*}{dt}(0) + (u^*(0) - f(0))\frac{h}{2} + O(h^2),$$

o ile u jest dostatecznie gładka. Zatem - korzystając z obu faktów - możemy skonstruować równanie różnicowe:

$$\partial_h u_h(0) - u_h(0)\frac{h}{2} = -f(0)\frac{h}{2}$$

przybliżające warunek Neumanna w punkcie $x = 0$ z wyższym rzędem.

Skonstruuj analogiczne równanie różnicowe przybliżające warunek Neumanna w punkcie $x = 1$ z wyższym rzędem. Pokaż, że rząd lokalnego błędu aproksymacji tego schematu względem parametru siatki h wynosi dwa w dyskretnej normie maksimum i dyskretnej normie L_h^2 dla odpowiednio gładkiego rozwiązania. Przetestuj w octavie rząd lokalnego błędu schematu w normie dyskretnej maksimum dla $u = \cos(\pi x)$ metodą połowienia kroków.

Ćwiczenie 8.10. Rozpatrzmy modelowe zadanie dwuwymiarowe na kole o średnicy jeden tzn. (7.8) dla $\overline{\Omega} = \overline{K}(0,1)$. Dobierzmy siatkę na płaszczyźnie o parametrze h równomierną zawierającą punkt $(0,0)$, tzn. $\{(k*h, l*h)\}_{k,l}$.

Za Ω_h uznajmy wszystkie punkty siatki, które należą do Ω i wszystkie punkty przecięcia prostych zadających siatkę z brzegiem Ω. Te punkty przecięcia uznajemy za brzegowe punkty siatki. Otrzymujemy oczywiście siatkę nierównomierną, bo odległość między brzegowym punktem siatki, a jego sąsiadem wewnętrznym jest mniejsza od h (poza ewentualnie pojedynczymi punktami).

Tu warunek brzegowy możemy zadać dokładnie. Pojawia się pytanie: jak przybliżyć drugą pochodną w punktach wewnętrznych siatki, których punkty sąsiednie są na brzegu?

Definiujemy w takim punkcie $x \in \Omega_h$ (załóżmy, że tylko jego prawy sąsiad $x_r = x + \tilde{h}\vec{e}_1$ jest na brzegu):

$$L_h u(x) = a * u(x - h\vec{e}_1) + bu(x) + cu(x + \tilde{h}\vec{e}_1) + -\partial\overline{\partial}_2 u(x).$$

dla pewnych parametrów a, b, c.

Jeśli x ma dwa punkty sąsiednie leżące na brzegu (powiedzmy prawy i dolny punkt sąsiedni), tzn. $x_p = x + \tilde{h}\vec{e}_1, x_d = x - \hat{h}\vec{e}_1$ są na brzegu, to oczywiście musimy wyznaczyć całe równanie różnicowe:

$$L_h u(x) = a * u(x - h\vec{e}_1) + bu(x) + cu(x + \tilde{h}\vec{e}_1) + \alpha * u(x - h\vec{e}_1) + \beta u(x) + \gamma u(x - \hat{h}\vec{e}_2).$$

Zadanie: Wyznacz odpowiednie parametry a, b, c, czy α, β, γ tak, aby lokalny błąd schematu $|L_h u(x) - Lu(x)|$ był możliwie mały, tzn. żeby schemat posiadał możliwie wysoki rząd lokalnego błędu aproksymacji względem h w dyskretnej normie maksimum.

9. Metoda różnic skończonych - stabilność schematów dla zadań eliptycznych w normie maksimum

W tym rozdziale zajmiemy się przedstawieniem metod badania stabilności schematów różnicowych dla zadań liniowych w dyskretnej normie maksimum.

Będziemy badali stabilność schematu zapisanego w formie (8.5)-(8.6). Dla $x \in \Omega_h$ możemy zapisać (8.5) jako:

$$L_h u(x) \equiv= \sum_{y \in N_h(x)} A(x,y) u_h(y) = f_h(x), \tag{9.1}$$

gdzie $N_h(x)$ jest podzbiorem $\overline{\Omega}_h$ punktów, dla których $A(x,y) \neq 0$, czyli uwzględnionych w równaniu dla tego x.

Jeśli $x \in \Gamma_{k,h}$, to dla (8.6) zachodzi:

$$l_{k,h} u(x) \equiv \sum_{y \in N_h(x)} A(x,y) u_h(y) = g_{k,h}(x),$$

gdzie $N_h(x)$ jest zdefiniowane analogicznie jak poprzednio. $N_h(x)$ jest zdefiniowane jednoznacznie.

$N_h(x)$ nazywamy otoczeniem siatkowym punktu x. Wprowadzimy również otoczenie siatkowe nakłute: $N'_h(x) = N_h(x) \setminus \{x\}$. Oczywiście $N_h(x)$ może być jednopunktowe, wtedy $N'_h(x)$ jest zbiorem pustym.

Zapiszmy schemat (8.5)-(8.6) jako:

$$\mathcal{L}_h u_h(x) \equiv \sum_{y \in N_h(x)} A(x,y) u_h(y) = \psi_h(x) \qquad x \in \overline{\Omega}_h, \tag{9.2}$$

gdzie

$$\psi_h(x) = \begin{cases} f_h(x) & x \in \Omega_h \\ g_{k,h}(x) & x \in \Gamma_{k,h} \quad s = 1, \ldots, s. \end{cases}$$

Wtedy zachodzi następujące twierdzenie, pozwalające na wykazanie stabilności niektórych schematów w normie dyskretnej maksimum:

Twierdzenie 9.1. *Niech $\mathcal{L}_h$ dla (8.5)-(8.6) będzie w formie (9.2). Załóżmy, że dla pewnej stałej $\alpha > 0$ i dla $h \leqslant h_0$:*

$$|A(x,x)| - \sum_{y \in N'_h(x)} |A(x,y)| \geqslant \alpha \qquad \forall x \in \overline{\Omega}_h.$$

Wtedy

$$\|u_h\|_{\infty,h} \leqslant \frac{1}{\alpha} \max(\|f_h\|_{\infty,h,\Omega_h} + \sum_{k=1}^{s} \|g_{k,h}\|_{\infty,h,\Gamma_{k,h}}).$$

Dowód. Widzimy, że $\|u_h\|_{\infty,h} = \max_{x\in\overline{\Omega}_h} |u_h(x)| = |u_h(x_0)|$ dla pewnego $x_0 \in \overline{\Omega}_h$.

Rozpatrzmy równanie ze schematu dla tego punktu:

$$\begin{aligned}
|\psi_h(x_0)| &= |\sum_{y\in N_h(x_0)} A(x_0,y)u_h(y)| \\
&\geqslant |A(x_0,x_0)||u_h(x_0)| - \sum_{y\in N_h'(x_0)} |A(x_0,y)||u_h(y)| \\
&\geqslant |A(x_0,x_0)||u_h(x_0)| - \sum_{y\in N_h'(x_0)} |A(x_0,y)||u_h(x_0)| \\
&= \left(|A(x_0,x_0)| - \sum_{y\in N_h'(x_0)} |A(x_0,y)|\right)|u_h(x_0)| \geqslant \alpha|u_h(x_0)|,
\end{aligned}$$

czyli

$$\|u_h\|_{\infty,h} = |u_h(x_0)| \leqslant \frac{1}{\alpha}\max_{x\in\overline{\Omega}_h}|\psi_h(x)|.$$

□

Jak widzimy, jest to proste kryterium. Sprawdźmy je na naszym modelowym zadaniu (7.5)-(7.6):

Przykład 9.1. Dla zadania (7.5)-(7.6) otrzymujemy następujący układ:

$$\begin{aligned}
-\frac{1}{h^2}u(x_{k-1}) + (\frac{2}{h^2}+c)u(x_k) - \frac{1}{h^2}u(x_{k+1}) &= f(x_k) \qquad k = 1,\ldots,N-1 \\
u(x_k) &= g(x_k) \qquad k = 0, N.
\end{aligned}$$

dla $x_k = a + k * h$.

Zatem $N_h(x_k) = \{x_{k-1}, x_k, x_{k+1}\}$ dla $k = 1,\ldots,N-1$ i $N_h(x) = \{x_k\}$ dla $x_k \in \{a,b\}$, tzn. dla $k = 0, N$. Sprawdzamy założenie twierdzenia:

$$|A(x_k,x_k)| - \sum_{y\in N_h(x_k)\setminus\{x_k\}} |A(x_k,y)| = \begin{cases} c & k = 1,\ldots,N-1 \\ 1 & k = 0, N. \end{cases}$$

Zatem $\alpha = \min\{1,c\}$ i z naszego kryterium, tzn. z twierdzenia 9.1, otrzymujemy stabilność zadania przybliżonego w normie dyskretnej supremum tylko w przypadku $c > 0$ ze stałą $\frac{1}{\alpha} = \max\{1, \frac{1}{c}\}$.

Powyższe oszacowanie sugeruje, że jeśli $c = 0$, to schemat może nie być stabilny w dyskretnej normie maksimum. Okaże się, że istnieją jednak inne kryteria badania stabilności, które są bardziej precyzyjne. Przedstawimy je poniżej.

9.1. Różnicowa zasada maksimum

Jak wiadomo, por. np. rozdział 6.4 w [11], dla równania eliptycznego spełnionych jest szereg zasad maksimum. Okaże się, że odpowiednio skonstruowane schematy różnicowe, czyli problemy przybliżone (różnicowe), spełniają analogiczne różnicowe zasady maksimum. Korzystając z tych zasad będziemy mogli wykazać stabilność tychże schematów.

Załóżmy, że operator $\mathcal{L}_h$ określony na $\overline{\Omega}_h$ jest w formie (9.2).

Definicja 9.1. Operator $\mathcal{L}_h$ w postaci (9.2) będziemy nazywać *operatorem dodatniego typu* (ang. *positive operator*) w $\overline{\Omega}_h$, jeśli dla dowolnego $x \in \overline{\Omega}_h$

1. $A(x,x) > 0$,

2. $A(x,y) < 0 \qquad \forall y \in N_h'(x)$,
3. $\sum_{y \in N_h(x)} A(x,y) \geqslant 0$.

Dodatkowo dla operatora typu dodatniego przedstawiamy siatkę Ω_h jako dwa rozłączne zbiory $\overline{\Omega}_h = \sum_{k=1,2} \Omega_h^{(1)} \cup \Omega_h^{(2)}$ zdefiniowane jako:

$$\Omega_h^{(1)} = \{x \in \overline{\Omega}_h : \sum_{y \in N_h(x)} A(x,y) = 0\}$$

i

$$\Omega_h^{(2)} = \overline{\Omega}_h \setminus \Omega_h^{(1)} = \{x \in \overline{\Omega}_h : \sum_{y \in N_h(x)} A(x,y) > 0\}.$$

Wprowadzamy jeszcze jedną definicję:

Definicja 9.2. Załóżmy, że $\mathcal{L}_h$ w postaci (9.2) jest operatorem dodatniego typu w $\overline{\Omega}_h$, dla którego zachodzi warunek: $\Omega_h^{(2)} \neq \emptyset$ i $\overline{\Omega}_h$ jest zbiorem skończonym. Wtedy powiemy, że $\overline{\Omega}_h$ spełnia *warunek spójności siatki* (ang. *mesh connectivity condition, mesh is connected*), jeśli dla dowolnego $x \in \Omega_h^{(1)}$ istnieje ciąg elementów siatki $\{x_i\}_{i=1}^N \subset \Omega_h^{(1)}$ i $y \in \Omega_h^{(2)}$ taki, że $x_1 = x$, i $x_{i+1} \in N_h(x_i)$ dla $i = 1, \ldots, N-1$ i $y \in N_h(x_N)$.

Wtedy zachodzi następująca różnicowa zasada maksimum:

Twierdzenie 9.2 (Różnicowa zasada maksimum - ang. *finite difference maximum principle*)**.** *Załóżmy, że $\mathcal{L}_h$ w postaci (9.2) jest operatorem dodatniego typu w $\overline{\Omega}_h$, i że $\overline{\Omega}_h$ jest zbiorem skończonym spełniającym warunek spójności siatki. Wtedy, jeśli*

$$\mathcal{L}_h u_h(x) \geqslant 0 \qquad \forall x \in \overline{\Omega}_h,$$

to

$$u_h(x) \geqslant 0 \qquad \forall x \in \overline{\Omega}_h.$$

Dowód można znaleźć w Rozdziale 10 w [10].

Wniosek 9.1. *Załóżmy, że spełnione są założenia twierdzenia 9.2. Wtedy zadanie (9.2) ma jednoznaczne rozwiązanie.*

Dowód. Załóżmy, że zadanie (9.2) ma dwa różne rozwiązania u_k dla $k = 1, 2$. Wtedy z twierdzenia 9.2 wynika, że $\mathcal{L}_h(u_1 - u_2) = 0$ zatem $(u_1 - u_2) \geqslant 0$ ale i $(u_2 - u_1) \geqslant 0$, czyli $u_1 = u_2$. Z kolei zauważmy, że (9.2) jest układem równań liniowych, więc jednoznaczność rozwiązania z prawą stroną równą zero jest równoważna istnieniu rozwiązania dla dowolnego ψ_h. □

Jako kolejny wniosek z różnicowej zasady maksimum otrzymujemy następujące kryterium porównawcze:

Twierdzenie 9.3. *Załóżmy, że spełnione są założenia twierdzenia 9.2 oraz niech*

$$\mathcal{L}_h u_h(x) = f_h(x) \qquad \mathcal{L}_h v_h(x) = g_h(x) \qquad x \in \overline{\Omega}_h.$$

Wtedy, jeśli

$$|f_h(x)| \leqslant g_h(x) \qquad x \in \overline{\Omega}_h,$$

to

$$|u_h(x)| \leqslant v_h(x) \qquad x \in \overline{\Omega}_h.$$

Dowód. Niech $z_h = u_h - v_h$, a $w_h = u_h + v_h$. Stąd

$$\mathcal{L}_h(-z_h(x)) = -f_h(x) + g_h(x) \geqslant 0, \qquad \mathcal{L}_h w_h(x) = f_h(x) + g_h(x) \geqslant 0 \qquad x \in \overline{\Omega}_h.$$

Zatem z twierdzenia 9.2 otrzymujemy:

$$z_h(x) \leqslant 0, \qquad w_h(x) \geqslant 0 \qquad x \in \overline{\Omega}_h,$$

a stąd otrzymujemy $|u_h(x)| \leqslant v_h(x)$ dla $x \in \overline{\Omega}_h$. □

Z ostatniego twierdzenia otrzymujemy następujące kryterium badania stabilności w dyskretnej normie maksimum:

Twierdzenie 9.4 (kryterium stabilności z różnicowej zasady maksimum)**.** *Załóżmy, że spełnione są założenia twierdzenia 9.2*
oraz, że istnieje nieujemna funkcja v_h określona na $\overline{\Omega}_h$ taka,że

$$0 \leqslant v_h \leqslant M, \qquad \mathcal{L}_h v_h \geqslant 1.$$

Wtedy u_h - rozwiązanie (9.2) z prawą stroną ψ_h, spełnia:

$$\|u_h\|_{\infty,h} = \max_{x \in \overline{\Omega}_h} |u_h(x)| \leqslant M \|\psi_h\|_{\infty,h}.$$

Dowód. Dla prostoty załóżmy, że $\|\psi_h\|_{\infty,h} = 1$ ($\mathcal{L}_h$ jest liniowe, więc zawsze możemy przeskalować u_h i ψ_h przez stałą różną od zera). Wtedy

$$|\mathcal{L}_h u_h(x)| = |\psi_h(x)| \leqslant 1 \leqslant \mathcal{L}_h v_h \qquad x \in \overline{\Omega}_h,$$

zatem z twierdzenia 9.3 otrzymujemy:

$$|u_h(x)| \leqslant v_h(x) \leqslant M = M\|\psi_h\|_{\infty,h} \qquad x \in \overline{\Omega}_h.$$

□

Przykład 9.2. Powróćmy do dyskretyzacji naszego modelowego zadania, tzn. do (7.5)-(7.6). Pozostawiamy jako proste zadanie sprawdzenie, że operator $\mathcal{L}_h$ w tym przypadku jest operatorem dodatniego typu, i że siatka spełnia warunek spójności.

Aby pokazać oszacowanie stabilności korzystając z naszego kryterium należy znaleźć funkcję nieujemną ψ określoną na $\overline{\Omega}$, czyli w szczególności na każdej siatce ograniczonej, taką że

$$\mathcal{L}_h\psi \geqslant 1.$$

Na brzegu widzimy, że $\mathcal{L}_h\psi(x) = \psi(x)$ dla $x \in \{a, b\}$, więc wystarczy przyjąć ψ takie, że $\psi \geqslant 1$ na brzegu Ω.

Najprościej będzie znaleźć funkcję ψ taką, że $L\psi = -\frac{d^2\psi}{dx^2} \geqslant 2$. Następnie, korzystając z tego, że rząd aproksymacji zadania przybliżonego jest dwa w każdym punkcie siatki, tzn.

$$|(L\psi - \mathcal{L}_h r_h\psi)(x)| = |(L\psi - L_h r_h\psi)(x)| = O(h^2) \qquad x \in \Omega_h,$$

możemy wywnioskować, że istnieje stała h_0 taka, że dla $h \leqslant h_0$ funkcja

$$\psi_h(x) = \psi(x) = r_h\psi(x) \qquad x \in \overline{\Omega}_h$$

spełnia $\mathcal{L}_h\psi_h \geqslant 1$.

W naszym przypadku np. dla $c \geqslant 0$ wystarczy zdefiniować:

$$\psi(x) = 1 + \left(\frac{b-a}{2}\right) + (x-a) * (x-b).$$

Wtedy $0 \leqslant \psi \leqslant 1 + \frac{b-a}{2} + \frac{(b-a)^2}{4} = M$ i $-\frac{d^2\psi}{dx^2} + c * \psi \geqslant -\frac{d^2\psi}{dx^2} = 2$. Zatem z naszego kryterium otrzymujemy dla $h \leqslant h_0$, że

$$\|u_h\|_{\infty,h} \leqslant \left(1 + \frac{b-a}{2} + \frac{(b-a)^2}{4}\right) \max\{|g(a)|, |g(b)|, \|f_h\|_{\infty,\Omega_h,h}\},$$

czyli stabilność w dyskretnej normie maksimum.

Proszę zauważyć, że stała w oszacowaniu *nie zależy* od stałej c, za to - inaczej niż w przypadku poprzedniego prostszego kryterium, zależy od długości odcinka (a, b).

Jeśli rozwiązanie (7.1) jest w $C^4([a, b])$, to otrzymujemy, że:

$$\|\mathcal{L}_h r_h u - Lu\|_{\infty,h,\Omega_h} = O(h^2)$$

dla $\mathcal{L}_h$ z (8.7), a warunki brzegowe spełnione są dokładnie. Zatem, korzystając z twierdzenia 8.1, otrzymujemy:

$$\|r_h u - u_h\|_{\infty,h} = O(h^2).$$

9.2. Zadania

Ćwiczenie 9.1. Zbadaj stabilność schematu (8.10) dyskretyzacji modelowego problemu dwuwymiarowego w dyskretnej normie maksimum dla $c > 0$.

Ćwiczenie 9.2. Sprawdź, czy operator z (8.10) jest dodatniego typu i zbadaj stabilność schematu (8.10) dyskretyzacji modelowego problemu dwuwymiarowego w dyskretnej normie maksimum dla $c = 0$ korzystając z różnicowej zasady maksimum.

Ćwiczenie 9.3. Rozpatrzmy problem $-\frac{du}{dt}(t) + c * u(t) = f(t)$ dla $t \in (0, 1)$ i f gładkiej funkcji z warunkiem brzegowym Neumanna $\frac{du}{dt}(s) = 0$ dla $s \in \{0, 1\}$ oraz schemat różnicowy na siatce jednorodnej $\overline{\Omega}_h = \{x_k\}_{k=0}^N$ dla $x_k = k * h$:

$$-\overline{\partial}\partial_h u_h(x_k) + cu_h(x_k) = f(x_k) \qquad k = 1, \ldots, N-1$$

z $\partial_h u_h(x_0) = \overline{\partial}_h u_h(x_N) = 0$. Czy to zadania wyjściowe oraz zadanie dyskretne mają jednoznaczne rozwiązanie dla $c > 0$?

Zbadaj rząd tego schematu oraz stabilność w dyskretnej normie maksimum dla stałej $c > 0$. Podaj oszacowanie błędu dyskretnego w dyskretnej normie maksimum w terminach $O(h^p)$.

Ćwiczenie 9.4. Zbadaj rząd i stabilność w normie maksimum schematu skonstruowanego analogicznie jak schemat (8.10) dyskretyzacji modelowego problemu dwuwymiarowego: $-\triangle u + c * u = f$ w $\Omega = (0,1)^2$ z zerowym warunkiem Dirichleta na brzegu kwadratu oprócz krawędzi $\Gamma_1 = 0 \times (0,1)$, gdzie jest postawiony zerowy warunek brzegowy Neumanna tzn. $u(s) = 0$ dla $s \in \partial\Omega \setminus \Gamma_1$ i $\frac{\partial u}{\partial n}(0,s) = -\frac{\partial u}{\partial x}(0,s) = 0$ na Γ_1.

Warunek brzegowy na Γ_1 przybliżamy w schemacie różnicowym przez odpowiednią różnicę skończoną wprzód, tzn. przez $\partial_1 u_h(0, k*h)$ dla $k = 1, \ldots, N-1$ z $h = \frac{1}{N}$.

Ćwiczenie 9.5. Zbadaj stabilność w dyskretnej normie maksimum schematu z ćwiczenia 8.6.

Ćwiczenie 9.6. Zbadaj stabilność w dyskretnej normie maksimum schematu z ćwiczenia 8.9.

10. Metoda różnic skończonych - stabilność schematów dla zadań eliptycznych w normach energetycznych

Materiał w poniższym rozdziale jest materiałem dodatkowym, tzn. nie wchodzi w zakres materiału przedstawianego na wykładzie.

10.1. Wprowadzenie - stabilność dla modelowego zadania

W tym rozdziale przedstawimy krótki zarys innej metody badania stabilności zadań przybliżonych otrzymanych za pomocą metody różnic skończonych, tym razem, w dyskretnej normie L^2_h. Jest to metoda analogiczna do metody badania stabilności zadań różniczkowych w równaniach fizyki matematycznej, por. [11].

Przedstawimy tę metodę teraz dla naszej modelowej dyskretyzacji (7.5) z jednorodnymi warunkami brzegowymi:

$$\begin{aligned} -\partial\overline{\partial}u_h(x) + c * u_h(x) = f(x) \qquad & x \in \Omega_h, \\ u(x) = 0 \qquad & x \in \partial\Omega_h. \end{aligned} \tag{10.1}$$

W przypadku niejednorodnych warunków brzegowych dla $c = 0$, zamiana zmiennych: $v(t) = u(t) + g(a) + \frac{g(b)-g(a)}{b-a}(t-a)$ dla u rozwiązania zadania z zerowymi warunkami brzegowymi daje v - rozwiązanie (7.5).

Proszę zauważyć, że dla tego zadania dyskretnego zachodzi też stabilność w dyskretnej normie maksimum, por. rozdział 9.

Przyjmujemy oznaczenie $\overline{\Omega}_h = \{x_k = a+k*h : k = 0, \ldots, N\}$. Wprowadzamy do przestrzeni $L^2_h(\overline{\Omega}_h)$ wszystkich funkcji określonych na siatce $\overline{\Omega}_h$ następujący iloczyn skalarny:

$$[u, v]_h = h \sum_{x\in\overline{\Omega}_h} u(x)v(x) = h\sum_{k=0}^{N} u_k v_k$$

będący dyskretnym odpowiednikiem iloczynu skalarnego typu $L^2(\Omega)$. Tutaj $u_k = u(x_k)$. Wprowadzamy dodatkowo oznaczenia:

$$[u, v)_h = h\sum_{k=0}^{N-1} u_k v_k, \qquad (u, v]_h = h\sum_{k=1}^{N} u_k v_k, \qquad (u, v)_h = h\sum_{k=1}^{N-1} u_k v_k.$$

Potrzebujemy następujących odpowiedników różnicowych wzorów na całkowanie przez części nazywanych: *różnicowymi wzorami na sumowanie przez części* (ang. *finite difference summing by parts formulas*):

$$\begin{aligned} h * \sum_{k=1}^{N-1} \partial u_k v_h &= -h * \sum_{k=1}^{N} u_k \overline{\partial} v_k + u_{N+1}v_{N+1} - u_1 v_0 \\ h * \sum_{k=1}^{N-1} \overline{\partial} u_k v_k &= -h * \sum_{k=0}^{N-1} u_k \partial v_h + u_N v_{N+1} - u_0 v_0. \end{aligned}$$

Tutaj $\overline{\partial} u_k = \overline{\partial} u(x_k) = h^{-1}(u_k - u_{k-1})$ i $\partial u_k = \partial u(x_k) = h^{-1}(u_{k+1} - u_k)$. Dowód tych wzorów pozostawiamy jako proste zadanie, por. ćwiczenie 10.1. Możemy je przedstawić z wykorzystaniem naszej notacji:

$$\begin{array}{rcl} (\partial u, v)_h & = & -(u, \overline{\partial} v]_h + u_{N+1} v_{N+1} - u_1 v_0, \\ (\overline{\partial} u, v)_h & = & -[\overline{\partial} u, v)_h + u_N v_{N+1} - u_0 v_0. \end{array} \tag{10.2}$$

Zauważmy, że $\partial\overline{\partial} u_k = \overline{\partial}\partial u_k$ dla $k = 1, \ldots, N-1$ zatem z powyższych wzorów dla u widzimy, że dla $u_0 = u_N = 0$:

$$(-\partial\overline{\partial} u, u)_h = (-\overline{\partial}\partial u, u)_h = [\partial u, \partial u)_h = (\overline{\partial} u, \overline{\partial} u]_h. \tag{10.3}$$

Prawdziwy jest również dyskretny odpowiednik nierówności Friedrichsa:

Twierdzenie 10.1 (różnicowa nierówność Friedrichsa)**.** *Dla $u \in L^2_h(\overline{\Omega}_h)$ takiej, że $u_0 = u_N = 0$ prawdziwa jest nierówność*

$$\|u\|^2_{0,h} \leqslant (b-a)^2 [\partial u, \partial u)_h = (b-a)^2 (\overline{\partial} u, \overline{\partial} u]_h.$$

Dowód pozostawiamy jako zadanie, por. ćwiczenie 10.1.

Weźmy $-\partial\overline{\partial} u_k$ dla u_h rozwiązania (10.1), przemnóżmy przez $h * u_k$ i zsumujmy po $k = 1, \ldots, N-1$. Wtedy, korzystając z wzorów na sumowanie przez części (10.2), otrzymujemy

$$(-\partial\overline{\partial} u_h, u_h)_h + c(u_h, u_h)_h = (\overline{\partial} u_h, \overline{\partial} u_h]_h + c(u_h, u_h)_h = (f_h, u_h)_h.$$

Możemy skorzystać z różnicowej nierówności Friedrichsa, por. twierdzenie 10.1:

$$\|u_h\|^2_{0,h} \leqslant (b-a)^2 (\overline{\partial} u_h, \overline{\partial} u_h]_h \leqslant (b-a)^2 (f_h, u_h)_h \leqslant (b-a)^2 \|f_h\|_{0,h,\Omega_h} \|u_h\|_{0,h},$$

a stąd otrzymujemy oszacowanie:

$$\|u_h\|_{0,h} \leqslant (b-a) \|f_h\|_{0,h,\Omega_h}.$$

W przypadku $c > 0$ otrzymujemy oszacowanie bez użycia nierówności Friedrichsa:

$$\|u_h\|_{0,h} \leqslant \sqrt{c^{-1}} \|f_h\|_{0,h,\Omega_h}.$$

Uzyskaliśmy stabilność w dyskretnej normie L^2_h, z której wynika też istnienie jednoznacznego rozwiązania równego zero dla $f_h = 0$. Stąd wynika istnienie jednoznacznego rozwiązania.

Weźmy $r_h u \in L^2_h(\overline{\Omega}_h)$ zdefiniowane jako $r_h u(x) = u(x)$ dla $x \in \overline{\Omega}_h$. Takie obcięcie jest zdefiniowane poprawnie dla dowolnej funkcji ciągłej. Zauważmy, że zbiór funkcji ciągłych na $\overline{\Omega}$ jest gęsty w $L^2(a,b)$. Dodatkowo

$$\|r_h u\|_{0,h} \to \|u\|_{L^2(a,b)} \qquad h \to 0$$

dla dowolnej funkcji ciągłej na $[a,b]$ oraz jeśli rozwiązanie (7.5) jest w $C^4([a,b])$, to

$$\|L_h r_h u - Lu\|_{0,h} = O(h^2).$$

Korzystając z twierdzenia 8.1 otrzymujemy:

$$\|r_h u - u_h\|_{0,h} = O(h^2). \tag{10.4}$$

Ten przykład jest prosty, ale w ten sam sposób można badać bardziej skomplikowane schematy różnicowe dla zadań postawionych w obszarach w dwóch czy więcej wymiarach.

10.2. Stabilności w normach energetycznych

Przedstawimy teraz ogólną teorię stabilności w dyskretnych normach energetycznych. Dyskretne normy energetyczne są analogiczne do tzw. norm energetycznych, w których bada się stabilność rozwiązań wyjściowych zadań różniczkowych z wykorzystaniem teorii równań fizyki matematycznej.

Zakładamy, że rozpatrujemy rodzinę skończenie wymiarowych przestrzeni Hilberta H_h z iloczynem skalarnym $(\cdot,\cdot)_h$ oraz operator $A_h : H_h \to H_h$. Interesuje nas zadanie dyskretne:

$$A_h u_h = f_h. \tag{10.5}$$

Powiemy, że operator liniowy $A : H_h \to H_h$ jest samosprzężony w H_h, jeśli $A = A^*$ dla $A^* : H_h \to H_h$ zdefiniowanego jako

$$(A^* u, v)_h = (u, Av)_h \qquad \forall u, v \in H_h.$$

Powiemy, że A jest dodatnio określony (nieujemnie określony), jeśli

$$(Au, u)_h > 0 \quad ((Au, v)_h \geqslant 0) \qquad \forall u \in H_h, \quad u \neq 0.$$

Nierówność operatorową $A > B$ $(A \geqslant B)$ definiujemy jako $A - B > 0$ $(A - B \geqslant 0)$. Zauważmy, że jeśli $A = A^* > 0$ to $(u, v)_A = (Au, v)_h$ jest poprawnie zdefiniowanym iloczynem skalarnym, który nazywamy iloczynem skalarnym energetycznym dla operatora A. Oznaczmy $\|u\|_A = (u, u)_A^{1/2}$ jako normę energetyczną dla A. Zauważmy, że A^{-1} też jest samosprzężony dodatnio określonym operatorem. Stabilność w odpowiednich normach dyskretnych typu L^2, czy normach energetycznych pozwala nam badać następujące twierdzenie:

Twierdzenie 10.2. *Niech $A : H_h \to H_h$ będzie liniowym operatorem w przestrzeni Hilberta skończenie wymiarowej H_h. Wtedy, dla u_h rozwiązania (10.5) zachodzi:*

— *jeśli $A \geqslant \alpha_1 I$, to*
$$\|u\|_h \leqslant \alpha_1^{-1} \|f\|_h,$$

— *jeśli $A = A^* \geqslant \alpha_2 I$, to*
$$\|u\|_A \leqslant \alpha_2^{-1/2} \|f\|_h,$$

— *jeśli $A \geqslant \alpha_3 B$ dla $B = B^* > 0$, to*
$$\|u\|_B \leqslant \alpha_3^{-1} \|f\|_{B^{-1}},$$

gdzie α_k dla $k = 1, 2, 3$ są stałymi dodatnimi.

Dowód pozostawiamy jako zadanie, por. twierdzenia 10.10 w [10].

Przykład 10.1. Zastosujmy powyższe twierdzenia do badania stabilności w przestrzeni Hilberta $L_h^2(\Omega_h)$ funkcji określonych na $\Omega_h = \{x_k\}_{k=1,\ldots,N-1}$ dla $x_k = a + k * h$ z iloczynem skalarnym $(u, v)_h = \sum_{k=1}^{N-1} u_k v_k$ dyskretyzacji (10.1). Bierzemy, jak powyżej, $u_k = u(x_k)$ dla $u \in L_h^2(\Omega_h)$ przy czym przyjmujemy, że $u_0 = u_N = 0$.

Pokażemy, że nasz powyższy dowód stabilności bazował na tym, że odpowiedni operator różnicowy jest dodatnio określony w tej przestrzeni.

Definiujemy $A_h, B_h : L_h^2(\Omega_h) \to L_h^2(\Omega_h)$ jako

$$\begin{aligned} B_h u(x) &= -\partial\overline{\partial} u(x) \qquad x \in \Omega_h, \\ A_h u(x) &= -\partial\overline{\partial} u(x) + c * u(x) \qquad x \in \Omega_h. \end{aligned}$$

Wtedy, przyjmując że $u_0 = u_N = v_0 = v_N = 0$, otrzymujemy jak powyżej (por. wzory na sumowanie przez części (10.2)):

$$(B_h u, v)_h = [\partial u, \partial v)_h = (u, B_h v)_h,$$

a następnie, z różnicowej nierówności Friedrichsa, por. twierdzenie 10.1, dla $u \neq 0$ widzimy, że

$$(B_h u, u)_h = [\partial u, \partial u)_h \geqslant \frac{1}{(b-a)^2}(u, u)_h > 0,$$

czyli $B_h \geqslant \frac{1}{(b-a)^2} I$. A z kolei $A_h = B_h + c * I \geqslant \left(c + \frac{1}{(b-a)^2}\right) * I$, czyli jest to operator dodatnio określony i samosprzężony i zachodzi $A_h \geqslant B_h$. Zatem, z pierwszego podpunktu twierdzenia 10.2 otrzymujemy:

$$\|u_h\|_h \leqslant \left(c + \frac{1}{(b-a)^2}\right)^{-1} \|f_h\|_h,$$

a z drugiego i trzeciego - odpowiednio:

$$\begin{aligned} \|u_h\|_{A_h} &\leqslant \left(c + \frac{1}{(b-a)^2}\right)^{-1/2} \|f_h\|_h, \\ \|u_h\|_{B_h} &\leqslant \|f_h\|_{B_h^{-1}}. \end{aligned}$$

Przykład 10.2. Rozpatrzmy następujący problem różniczkowy, powstały z naszego modelowego problemu poprzez dodanie członu z pierwszą pochodną:

$$-u''(x) + b * u'(x) + c * u = f, \qquad u(0) = u(L) = 0$$

dla b, c stałych, przy czym $c \geqslant 0$. Dyskretyzujemy ten problem na siatce $\overline{\Omega}_h = \{x_k\}_{k=0,\ldots,N}$ dla $x_k = k * h$ dla $h = L/N$ w następujący sposób:

$$\begin{aligned} L_h u_h(x) = -\partial\overline{\partial} u_h(x) + b * \tilde{\partial} u + c * u_h(x) = f(x) \qquad & x \in \Omega_h, \\ u_h(0) = u_h(L) = 0 \qquad & x \in \partial\Omega_h. \end{aligned} \tag{10.6}$$

Tutaj

$$\tilde{\partial} u(x) = \frac{u(x+h) - u(x-h)}{2 * h}$$

jest ilorazem różnicowym centralnym. Zauważmy, że $\tilde{\partial} = 0.5(\partial + \overline{\partial})$. Można pokazać, że jeśli rozwiązanie $u \in C^4([0, L])$, to:

$$|L_h u(x) - f(x)| = O(h^2) \qquad x \in \overline{\Omega}_h,$$

co pozostawiamy jako zadanie. Z tego możemy wywnioskować, że rząd aproksymacji wynosi dwa, zarówno w normie dyskretnej maksimum, jak i w L_h^2.

Weźmy przestrzeń H_h z tym samym iloczynem skalarnym i operator B_h z przykładu 10.1.

Wtedy, z wzorów na różnicowe sumowanie przez części (10.2), otrzymujemy:

$$(\tilde{\partial} u, u)_h = 0.5 * (\partial u + \overline{\partial} u, u)_h = -0.5 * (u, \partial u + \overline{\partial} u)_h.$$

Stąd $(\tilde{\partial} u, u)_h = 0$. Zatem, choć L_h nie jest symetryczny (o ile $b \neq 0$), to jest operatorem dodatnio określonym i zachodzi:

$$(L_h u, u)_h = ((B_h + c * I)u, u)_h \geqslant \left(c + \frac{1}{L^2}\right) * (u, u)_h.$$

czyli $L_h \geqslant B_h + c * I \geqslant (c + \frac{1}{L^2}) * I$.

Z powyższego oszacowania możemy pokazać stabilność w normie $\|\cdot\|_{0,h}$ jak w przykładzie 10.1, a w konsekwencji zbieżność dyskretną z rzędem dwa, co pozostawiamy jako zadanie.

10.3. Zadania

Ćwiczenie 10.1. Udowodnij wzory na sumowanie przez części, tzn. (10.2) oraz różnicową nierówność Friedrichsa, tzn. twierdzenie 10.1.

Ćwiczenie 10.2. Zbadaj rząd i stabilność schematu z przykładu 10.2 dyskretyzacji modelowego problemu jednowymiarowego w $\|\cdot\|_{0,h}$ dla $c > 0$ i $c = 0$. Wykaż zbieżności z rzędem dwa w normie $\|\cdot\|_{0,h}$, o ile rozwiązanie wyjściowego problemu jest klasy C^4.

Ćwiczenie 10.3. Zbadaj stabilność schematu (8.10) dyskretyzacji modelowego problemu dwuwymiarowego w dyskretnej normie L^2 dla $c \geqslant 0$.

Ćwiczenie 10.4. Rozpatrzmy równanie różniczkowe na kwadracie $\Omega = (0,1)^2$: chcemy znaleźć $u \in C^2(\Omega) \cap C(\overline{\Omega})$:

$$-\triangle u + b_1 u_x + b_2 u_y + c * u = f \qquad \text{w} \quad [0,1]^2$$

z zerowym warunkiem brzegowym. Tu c, b_1, b_2 są stałymi, a c jest dodatkowo nieujemna.

Analogicznie do przykładu 10.2 i dyskretyzacji (8.10), skonstruuj schemat różnicowy wykorzystując odpowiednie pochodne centralne do aproksymacji pochodnych u_x, u_y.

Zbadaj rząd schematu i stabilność w w dyskretnej normie L^2.

Wskazówka. Postępuj analogicznie jak w przykładzie 10.2.

11. Metoda elementu skończonego - wprowadzenie

W tym rozdziale przedstawimy główne idee metody elementu skończonego na przykładzie modelowego zadania eliptycznego rzędu dwa na obszarze jednowymiarowym. Metoda elementu skończonego jest bardziej ogólna od metody różnic skończonych nawet dla zadań różniczkowych zadanych na obszarze w jednym wymiarze. Np. konstrukcje zadań przybliżonych dla warunków brzegowych różnego typu są dużo prostsze niż w przypadku metody różnic skończonych.

11.1. Metoda elementu skończonego dla modelowego zadania eliptycznego w jednym wymiarze

11.1.1. Słabe sformułowanie

Rozpatrzmy modelowe zadanie jednowymiarowe (7.1) z zerowymi warunkami brzegowymi i $c = 0$, którego rozwiązanie oznaczymy u_*.

Następnie weźmy dowolną funkcję ciągłą ϕ, która jest kawałkami C^1 na odcinku, tzn. która ma ciągłą pochodną poza skończoną ilością punktów taką, że $\phi(a) = \phi(b) = 0$. Przemnóżmy równanie $Lu = f$ przez tę funkcję. Ze wzoru na całkowanie przez części otrzymujemy:

$$\int_a^b -\frac{d^2u_*}{dx^2}\phi \, dx = \int_a^b \frac{du_*}{dx}\frac{d\phi}{dx} \, dx = \int_a^b f\phi \, dx.$$

Oczywiście tutaj $\frac{d\phi}{dx}$ jest zdefiniowana poza skończoną ilością punktów nieciągłości.

Zamiast zadania (7.1) możemy rozpatrzyć zadanie znalezienia funkcji u_* w odpowiedniej przestrzeni V funkcji określonych na odcinku $[a, b]$ zawierających funkcje kawałkami C^1 i zerujące się w końcach odcinka (na razie nie ustalajmy precyzyjnie o jaką przestrzeń chodzi) takiej, żeby

$$\int_a^b \frac{du_*}{dx}\frac{d\phi}{dx} \, dx = \int_a^b f\phi \, dx \quad \forall \phi \in V. \tag{11.1}$$

Oczywiście rozwiązanie (7.1) spełnia (11.1), a przy odpowiednim doborze V (oraz dostatecznej gładkości f) można pokazać, że rozwiązanie (11.1) również spełnia (7.1). Zadanie (11.1) nazywamy sformułowaniem uogólnionym (słabym, wariacyjnym) zadania (7.1). Zaś (7.1) nazywamy sformułowaniem klasycznym. Metoda elementu skończonego, która jest szczególnym przypadkiem metody Galerkina polega na tym, że wprowadzamy w specjalny sposób skończenie wymiarową podprzestrzeń przestrzeni V. Następnie szukamy rozwiązania w tej przestrzeni dyskretnej, które spełnia zadanie wariacyjne (11.1) z tym, że przestrzeń V jest zastąpiona przez naszą dyskretną podprzestrzeń.

Proszę zauważyć, że podejście wariacyjne jest inne od metody różnic skończonych, w której konstrukcja rozwiązania określonego na zbiorze dyskretnym (siatce) polega na zastąpieniu odpowiednich pochodnych w równaniu różniczkowym odpowiednimi ilorazami różnicowymi na tej siatce.

11.1.2. Element liniowy

Wprowadźmy podział (triangulacje) odcinka $[a,b]$ na pododcinki (elementy) $T_h([a,b]) = \{\tau_k\}$, gdzie $\tau_k = (x_k, x_{k+1})$ dla $a = x_0 < \ldots < x_{N-1} < x_N = b$. Za parametr tego podziału przyjmijmy $h = \max_k |x_k - x_{k-1}|$, a punkty x_k nazwiemy punktami nodalnymi (węzłami) tego podziału.

Oczywiście najprostszym podziałem jest podział równomierny, jeśli bierzemy $x_k = k * h$ dla $h = (b-a)/N$.

Zakładamy, że rozpatrujemy rodzinę podziałów z h dążącym do zera.

Teraz na bazie danego podziału $T_h([a,b])$ możemy wprowadzić przestrzeń dyskretną:

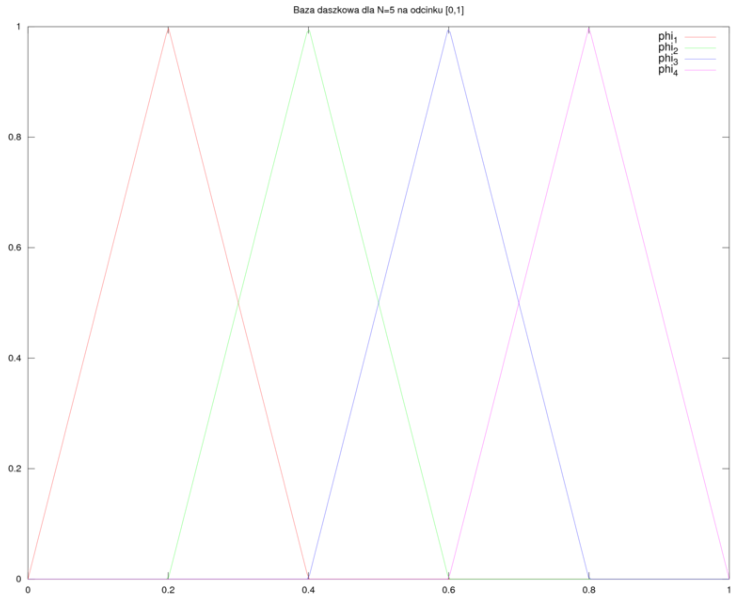
$$V^h = \{u \in C([a,b]) : u_{|\tau} \in P_1 \quad \forall \tau \in T_h([a,b]), \quad u(a) = u(b) = 0\},$$

gdzie $P_1 = span(1,x)$ jest przestrzenią wielomianów liniowych, tzn. stopnia nie większego niż jeden. Funkcje z tej przestrzeni to funkcje kawałkami liniowe (czyli klasyczne splajny liniowe) z zerowymi warunkami brzegowymi na brzegu, czyli w szczególności są to funkcje ciągłe kawałkami klasy C^1, zatem $V^h \subset V$.

Zadanie dyskretne polega na znalezieniu $u_h \in V^h$ takiego, że

$$\int_a^b \frac{du_h}{dx}\frac{d\phi}{dx}\,dx = \int_a^b f\phi\,dx \quad \forall \phi \in V^h. \tag{11.2}$$

Pozostawimy jako zadanie wykazanie istnienia jednoznacznego rozwiązania (11.2). Na razie



Rysunek 11.1. Baza daszkowa dla $N = 5$.

wprowadźmy tzw. funkcję nodalną: niech $\phi_k \in V^h$ dla $k \in \{1, \ldots, N-1\}$ będzie taką funkcją, że $\phi_k(x_k) = 1$ i $\phi_k(x_j) = 0$ dla $j \neq k$. Możemy podać wzór na taką funkcję:

$$\phi(x) = \begin{cases} 0 & x \notin [x_{k-1}, x_{k+1}], \\ \frac{x - x_{k-1}}{x_k - x_{k-1}} & x \in [x_{k-1}, x_k], \\ 1 - \frac{x - x_k}{x_{k+1} - x_k} & x \in [x_k, x_{k+1}]. \end{cases} \tag{11.3}$$

Widzimy wykres kilku takich funkcji, por. rysunek 11.1.

Nietrudno pokazać, że ϕ_k jest elementem V^h, i że $(\phi_1, \ldots, \phi_{N-1})$ tworzy bazę V^h taką, że jeśli $u \in V^h$, to

$$u = \sum_{k=1}^{N-1} u(x_k)\phi_k,$$

gdzie $u(x_k)$ wartość funkcji u w punkcie nodalnym x_k. Wstawiając $u_h = \sum_{k=1}^{N-1} u_h(x_k)\phi_k$ do (11.2) otrzymujemy następujący układ równań liniowych:

$$A_h \vec{u} = \vec{f} \tag{11.4}$$

z $\vec{u} = (u(x_1), \ldots, u(x_{N-1}))^T$,

$$A_h = (a_{kl})_{k,l=1}^{N-1}, \qquad a_{kl} = \int_a^b \frac{d\phi_k}{dx}\frac{d\phi_l}{dx}\,dx, \qquad \vec{f} = (f_1, \ldots, f_{N-1})^T$$

dla $f_k = \int_a^b f\phi_k\,dx$.

Zauważmy, że macierz A_h jest symetryczna i trójdiagonalna. Można wykazać, że jest dodatnio określona, więc można powyższy układ rozwiązać metodą przeganiania lub odpowiednim wariantem metody Choleskiego kosztem rzędu $c * N$ dla stałej c niezależnej od N.

11.1.3. Zbieżność

Zastanówmy się nad zbieżnością u_h do rozwiązania u_*. Najpierw trzeba ustalić w jakiej normie chcemy wykazać zbieżność.

Naturalną normą jest norma energetyczna związana z formą dwuliniową w słabym sformułowaniu (11.1):

$$\|u\|_V := \sqrt{\int_a^b \left|\frac{du}{dx}\right|^2 dx}.$$

Wprowadzając oznaczenie $e_h = u_* - u_h$ i odejmując (11.2) od (11.1) otrzymujemy:

$$\int_a^b \frac{de_h}{dx}\frac{d\phi}{dx}\,dx = \int_a^b \frac{d}{dx}(u_* - u_h)\frac{d\phi}{dx}\,dx = 0 \qquad \forall \phi \in V^h.$$

Następnie dla dowolnego $v_h \in V^h$ widzimy, że

$$\begin{aligned}\int_a^b \left|\frac{de_h}{dx}\right|^2 dx &= \int_a^b \frac{de_h}{dx}\frac{du_*}{dx}\,dx = \int_a^b \frac{de_h}{dx}\frac{du_*}{dx}\,dx - \int_a^b \frac{de_h}{dx}\frac{dv_h}{dx}\,dx \\ &= \int_a^b \frac{de_h}{dx}\frac{d}{dx}(u_* - v_h)\,dx.\end{aligned}$$

Korzystając z nierówności Schwarza w $L^2(a,b)$ otrzymujemy:

$$\|e_h\|_V^2 \leqslant \|e_h\|_V \|u_* - v_h\|_V,$$

czyli

$$\|e_h\|_V \leqslant \|u_* - v_h\|_V \qquad \forall v_h \in V^h.$$

Przyjmując za v_h liniowy interpolator u_* w punktach nodalnych, tzn. $I_h u_* := \sum_k u_*(x_k)\phi_k$, można wykazać:

$$\left\|\frac{d^k}{dx^k}(u_* - I_h u_*)\right\|_{\infty,[a,b]} \leqslant C_k h^{2-k} \left\|\frac{d^2 u_*}{dx^2}\right\|_{\infty,[a,b]} \qquad k = 0, 1 \tag{11.5}$$

dla pewnych stałych C_k niezależnych od h, u_* i $[a,b]$. Dowód tego oszacowania pozostawiamy jako zadanie, wynika on wprost z oszacowań błędu interpolacji dla splajnów liniowych.

Wtedy od razu otrzymujemy:

$$\|e_h\|_V \leqslant \|u_* - I_h u_*\|_V \leqslant C_1 h \left\| \frac{d^2 u_*}{dx^2} \right\|_{\infty,[a,b]},$$

czyli dla funkcji klasy C^2 błąd w normie energetycznej zachowuje się jak $O(h)$.

Można też wykazać, że w normie $L^2(a,b)$ zachodzi:

$$\|e_h\|_{L^2(a,b)} \leqslant C h^2 \left\| \frac{d^2 u_*}{dx^2} \right\|_{\infty,[a,b]},$$

czyli błąd zachowuje się jako $O(h^2)$ - co nie jest oczywiste.

Porównajmy to oszacowanie z oszacowaniem błędu z metody różnic skończonych (MRS) na siatce równomiernej dla tego samego zadania różniczkowego. Można pokazać wtedy zbieżność dyskretną $O(h^2)$ w dyskretnej normie L_h^2, por. (10.4), która w przybliżeniu odpowiada normie L^2, czyli możemy powiedzieć, że szybkość zbieżności w tym przypadku metody elementu skończonego i metody różnic skończonych jest tego samego rzędu. Ale w MRS musieliśmy założyć równomierność siatki i wyższą gładkość rozwiązania ($u_* \in C^4((a,b))$).

11.1.4. Inne przestrzenie elementu skończonego

Przestrzeń dyskretną V^h można też zdefiniować inaczej.

Dla danego podziału $T_h([a,b])$ zdefiniujemy następujące przestrzenie elementu skończonego dla dowolnego $p = 1, 2, 3, 4, \ldots$:

$$V_p^h = \{u \in C([a,b]) : u_{|\tau} \in P_p \quad \forall \tau \in T_h([a,b]), \quad u(a) = u(b) = 0\}$$

gdzie P_p jest przestrzenią wielomianów stopnia nie przekraczającego p.

Widzimy, że $V^h = V_1^h$. Przestrzeń V_2^h nazywamy przestrzenią elementu skończonego funkcji ciągłych kawałkami kwadratowych, a przestrzeń V_3^h - przestrzenią elementu skończonego funkcji ciągłych kawałkami kubicznych, czy inaczej - metodą elementu skończonego typu Lagrange'a kwadratową lub kubiczną. Możemy teraz postawić zadanie dyskretne, jak poprzednio, tzn. szukamy $u_{h,p} \in V_p^h$ takiego, że

$$\int_a^b \frac{du_{h,p}}{dx} \frac{d\phi}{dx}\, dx = \int_a^b f\phi\, dx \quad \forall \phi \in V_p^h. \tag{11.6}$$

Zadanie to ma jednoznaczne rozwiązanie. Analogicznie jak dla elementu liniowego możemy wprowadzić tu tzw. bazę nodalną w V_p^h. Wprowadzamy dodatkowe punkty wewnątrz odcinka $[x_k, x_{k+1})$ dla $k = 0, \ldots, N-1$:

$$x_{k,j} = x_k + \frac{j}{p}(x_{k+1} - x_k), \quad j = 0, \ldots, p-1.$$

Oczywiście $x_{k,0} = x_k$.

Dla każdego punktu $x_{k,j}$ oprócz $x_{0,0} = x_0 = a$ wprowadzamy funkcję bazową $\phi_{k,j} \in V_p^h$ taką, że

$$\phi_{k,j}(x_{l,i}) = \begin{cases} 1 & l = j \quad i = j \\ 0 & \text{w przeciwnym przypadku} \end{cases} \tag{11.7}$$

Można pokazać, że $(\phi_{k,j})_{k,j}$ jest bazą i to taką, że

$$\operatorname{supp}(\phi_{k,j}) = \begin{cases} [x_{k-1}, x_{k+1}] & j = 0, k > 0 \\ [x_k, x_{k+1}] & j \neq 0. \end{cases}$$

Powstaje pytanie: po co stosować przestrzenie V_p^h dla $p > 1$?

Jeśli rozwiązanie u_* jest bardziej regularne, tzn. należy do $C^{p+1}([a,b])$, to można wykazać, że dla $I_{h,p} = \sum_{k,j} u(x_{k,j})\phi_{k,j} \in V_p^h$:

$$\left\| \frac{d^k}{dx^k}(u_* - I_{h,p}u_*) \right\|_{\infty,[a,b]} \leqslant C_k h^{p+1-k} \left\| \frac{d^{p+1}u_*}{dx^{p+1}} \right\|_{\infty,[a,b]} \qquad k = 0, 1 \tag{11.8}$$

i stąd, jak poprzednio dla elementu liniowego, otrzymujemy, że

$$\|u_* - u_{h,p}\|_V \leqslant C_p h^p \left\| \frac{d^{p+1}u_*}{dx^{p+1}} \right\|_{\infty,[a,b]}, \tag{11.9}$$

czyli błąd zachowuje się jak $O(h^p)$ co oznacza, że w tej normie zachodzi zbieżność rzędu p.

11.2. Zadania

Ćwiczenie 11.1. Pokaż, że ϕ_k zdefiniowana w (11.3) jest w V^h, i że $(\phi_1, \ldots, \phi_{N-1})$ tworzy bazę V^h.

Ćwiczenie 11.2. Wykaż (11.4). Policz wszystkie różne od zera elementy macierzy A_h układu (11.4). Pokaż, że dla $c = 0$ i równomiernego podziału odcinka tzn. $x_k = a + k * h$ macierz ta jest równa macierzy dyskretyzacji metodą różnic skończonych dla tego samego zadania, pomnożonej przez parametr h, tzn. jest macierzą układu równań liniowych (7.7). Czy oba układy równań liniowych po przeskalowaniu przez h (7.7) są wtedy identyczne?

Rozwiązanie. Układy nie są identyczne, ponieważ prawe strony mogą być różne, jakkolwiek wtedy prawą stronę (7.7) możemy uznać za prostą aproksymację całek z prawej strony (11.4).

Ćwiczenie 11.3. Pokaż, że macierz A_h w (11.4) jest zawsze trójdiagonalna, symetryczna i dodatnio określona.

Wskazówka. Pokaż, że dla dowolnych funkcji $u = \sum_k u_k\phi_k, v = \sum_k v_k\phi_k \in V^h$ *zachodzi* $\vec{u}^T A_h \vec{v} = \int_a^b \frac{du}{dx}\frac{dv}{dx}\,dx$, *dla* $\vec{u} = (u_1, \ldots, u_{N-1})^T, \vec{v} = (v_1, \ldots, v_{N-1})^T$.

Ćwiczenie 11.4. Zaproponuj metodę rozwiązywania układu równań (11.4) będącą odpowiednią wersją metody eliminacji Gaussa dla macierzy symetrycznej trójdiagonalnej dodatnio określonej, której koszt wynosi $O(N)$.

Ćwiczenie 11.5 (laboratoryjne)**.** Dla podziału równomiernego na odcinku $[-1, 1]$ rozwiąż w octavie (11.4) dla znanego rozwiązania $u(x) = \sin(\pi * x)$, czyli dla $f = -\pi^2 * \sin(x)$. Prawą stronę możemy policzyć odpowiednią funkcją octave'a. Policz rozwiązania dyskretne dla $2h$ i h. Następnie policz normy dyskretne maksimum dla błędów $u_* - u_h$ i $u_* - u_{2h}$ (czyli maksima błędów w punktach nodalnych) i ich stosunek. (Zakładając, że błąd dla h wynosi $O(h^p)$, stosunek ten powinien w przybliżeniu wynosić 2^p).

Ćwiczenie 11.6 (częściowo laboratoryjne)**.** Udowodnij, że $\phi_{k,j}$ z (11.7) są w V_p^h, i że stanowią bazę tej przestrzeni. Wyprowadź bezpośrednie wzory na $\phi_{k,j}$. Narysuj w octave wykresy wszystkich $\phi_{k,j}$ na odcinku $[x_k, x_{k+1}]$ dla $[a, b] = [0, 2]$, $h = 1$ i $p = 2, 3$ przy użyciu funkcji octave'a **plot**().

Ćwiczenie 11.7. Korzystając z (11.8) wykaż (11.9) tzn., że

$$\|u_* - u_{h,p}\|_V = O(h^p)$$

o ile u_* jest dostatecznie gładka.

Wskazówka. Dowód przebiega identycznie jak w przypadku $p = 1$, tzn. liniowego elementu skończonego.

Ćwiczenie 11.8. Udowodnij jednowymiarową nierówność Friedrichsa, a mianowicie, że jeśli f jest funkcją ciągłą kawałkami klasy C^1 na $[a, b]$ i $u(a) = 0$, to

$$\int_a^b |f|^2 \, dx \leqslant (b-a)^2 \int_a^b \left|\frac{df}{dx}\right|^2 \, dx.$$

Ćwiczenie 11.9. Pokaż, że $(u, v)_V = \int_a^b \frac{du}{dx}\frac{dv}{dx}\, dx$ jest iloczynem skalarnym na V_p^h i - ogólniej na dowolnej przestrzeni funkcyjnej zawartej w przestrzeni funkcji ciągłych kawałkami C^1 zerujących się w końcach odcinka $[a, b]$. W szczególności $\|u\|_V$ jest normą na V_p^h.

Ćwiczenie 11.10. Wyprowadź układ równań liniowych

$$A_{h,p}\vec{u} = \vec{f},$$

którego rozwiązaniem jest wektor współczynników w bazie $\{\phi_{k,j}\}$, por. (11.7), rozwiązania $u_{h,p}$ zadania (11.6). Określ ilość elementów różnych od zera w macierzy, czy przy odpowiednim porządku indeksów funkcji bazy ta macierz może być jest pasmowa? Jeśli tak, to znajdź wielkość pasma. Czy jest symetryczna i dodatnio określona? Zaproponuj algorytm bezpośredni rozwiązywania tego układu kosztem $O(n)$, dla n wymiaru V_p^h.

Ćwiczenie 11.11. Udowodnij (11.5).

Wskazówka. Oszacowanie dla $k = 0$ wprost wynika z oszacowania błędu interpolacji Lagrange'a, por. np. [14], [18] lub w języku angielskim [17] lub [25]. Natomiast w przypadku $k = 1$ wystarczy zauważyć, że w każdym przedziale (x_k, x_{k+1}) dla $0 \leqslant k < N$ istnieje punkt ξ_k taki, że $u_*(\xi_k) = I_h u_*(\xi_k)$, a następnie skorzystać z twierdzenia o wartości średniej.

Ćwiczenie 11.12. Udowodnij (11.8).

Wskazówka. Ponownie jak w poprzednim zadaniu oszacowanie dla $k = 0$ wprost wynika z oszacowania błędu interpolacji Lagrange'a, por. np. [14], [18] lub w języku angielskim [17] lub [25].

12. Metoda elementu skończonego - wprowadzenia cd. Przypadek dwuwymiarowy.

W tym rozdziale przedstawimy kilka możliwie prostych dyskretyzacji skonstruowanych za pomocą metody elementu skończonego na przykładzie modelowego zadania eliptycznego na kwadracie.

12.1. Metoda elementu skończonego na kwadracie jednostkowym

Rozpatrzmy modelowe zadanie (7.8) z rozdziału 7.2, którego rozwiązanie oznaczymy przez u_*.

Słabe sformułowanie zadania (7.8): chcemy znaleźć $u_* \in V$ takie, że

$$\int_\Omega \nabla u_* \nabla v + cu\, v\, dx = \int_\Omega f v\, dx \qquad \forall v \in V. \tag{12.1}$$

Tutaj V jest odpowiednio dobraną przestrzenią funkcji ciągłych zerujących się na brzegu, dla których obie strony słabego sformułowania mają sens. Np. możemy wziąć domknięcie w odpowiedniej normie przestrzeni funkcji ciągłych zerujących się na brzegu, których słaba pochodna (pochodna w sensie dystrybucyjnym) jest w $L^2(\Omega)$. Później precyzyjniej ustalimy o jaką przestrzeń chodzi.

Jeśli problem (12.1) ma rozwiązanie dostatecznie gładkie, tzn. u_* posiada ciągłe pierwsze i drugie pochodne cząstkowe, to u_* jest rozwiązaniem wyjściowego zadania. Może się zdarzyć, że istnieje rozwiązanie (12.1), które nie jest nawet ciągłe.

12.1.1. Triangulacja obszaru

Wprowadźmy podział (triangulacje) kwadratu $\overline{\Omega} = [0,1]^2$ na trójkąty $T_h([0,1]^2) = \{\tau_k\}_k$ o jednakowym kształcie i wielkości. Najprościej jest podzielić kwadrat na równe kwadraty $([x^{(k)}, x^{(k+1)}] \times [x^{(l)}, x^{(l+1)}]$ dla $k,l = 0,\ldots,N-1$ i $x^{(k)} = k * h$ dla $h = 1/N$. Następnie każdy kwadrat dzielimy na dwa trójkąty prowadząc przekątną np. z lewego górnego rogu do dolnego prawego, por. rysunek 12.1.

Zauważmy, że $\bigcup_{\tau\in T_h(\Omega)} \overline{\tau} = \overline{\Omega}$ i $\partial\tau_k \cap \partial\tau_j$ jest zbiorem pustym, krawędzią lub wspólnym wierzchołkiem dla dowolnych różnych elementów tej triangulacji.

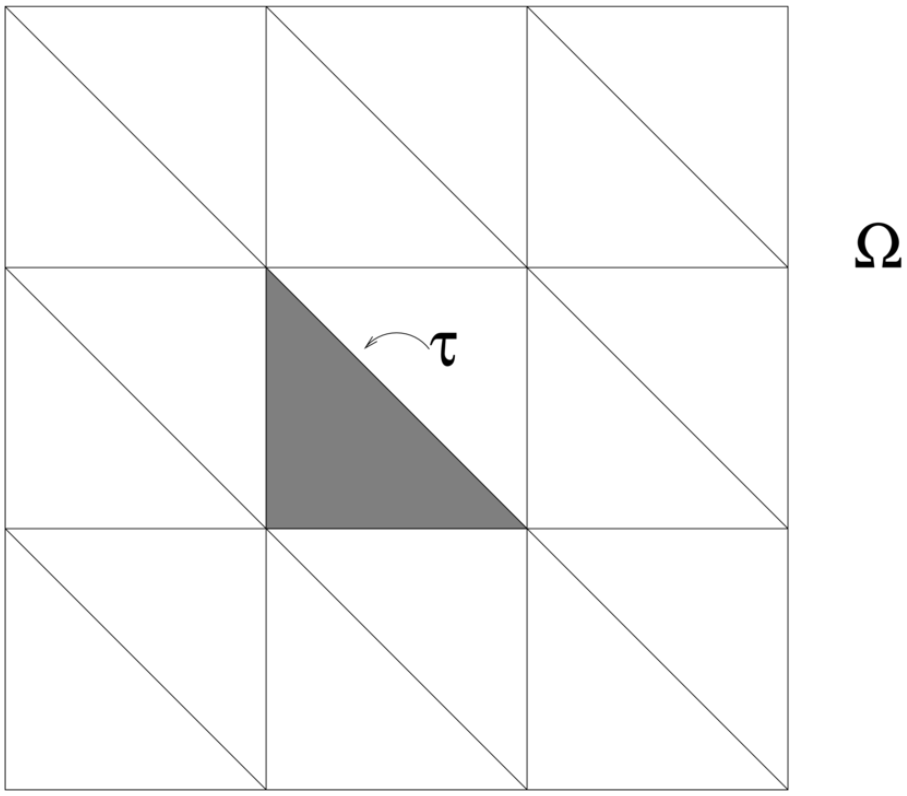
Za parametr tej dyskretyzacji przyjmujemy h, a punkty $x^{(kl)} = (x^{(k)}, x^{(l)})$ nazwiemy punktami nodalnymi tej triangulacji. Proszę zauważyć, że to nie jest tylko jedna możliwa triangulacja kwadratu. Możemy wybierać trójkąty na wiele sposobów tak, aby tylko zachowane zostały warunki, że trójkąty są rozłączne, suma ich domknięć tworzy cały kwadrat, ich wspólne części brzegów to wierzchołek, wspólna krawędź lub zbiór pusty.

Rozpatrujemy w ogólności rodzinę takich triangulacji z $h \to 0$.

12.1.2. Element liniowy

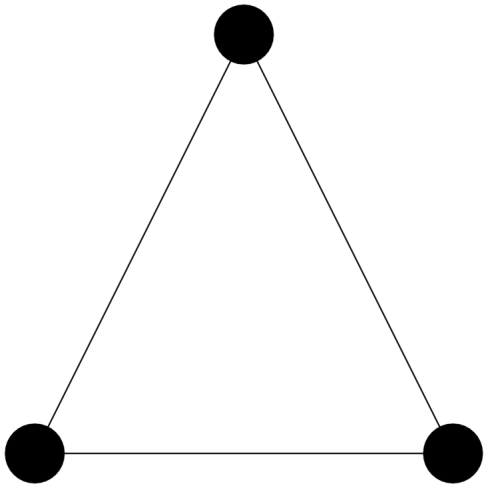
Możemy wprowadzić przestrzeń dyskretną:

$$V^h = \{u \in C(\overline{\Omega}) : u_{|\tau} \in P_1(\tau) \quad \forall \tau \in T_h(\Omega), \quad u(s) = 0 \quad \forall s \in \partial\Omega\},$$



Rysunek 12.1. Modelowa triangulacja złożona z trójkątów τ na kwadracie Ω.

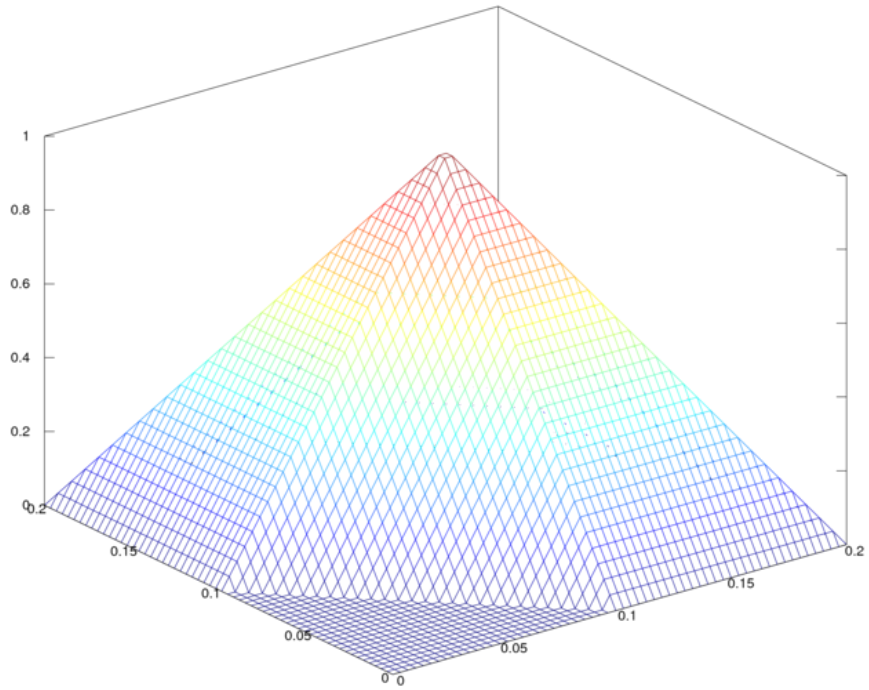
dla $P_1(\tau) = \text{span}(1, x_1, x_2)$ przestrzeni wielomianów liniowych na τ. Wielomian liniowy na trójkącie jest zdefiniowany poprzez wartości w wierzchołkach tego trójkąta, które nazywamy punktami swobody tego elementu skończonego, por. rysunek 12.2.



Rysunek 12.2. Wierzchołki trójkąta τ - punkty nodalne wielomianu liniowego na tym trójkącie.

Analogicznie jak w przypadku jednowymiarowym, możemy wprowadzić funkcje nodalne: definiujemy $\phi_{k,l} \in V^h$ dla $k, l \in \{1, \ldots, N-1\}$ jako funkcję, która spełnia $\phi_{k,l}(x^{(kl)}) = 1$ i $\phi_{k,l}(x^{(ij)}) = 0$ dla $x^{(kl)} \neq x^{(ij)}$ dla $x^{(kl)} = (x^{(k)}, x^{(l)})$.

W przypadku naszej prostej regularnej triangulacji możemy wyznaczyć wzory na te funkcje



Rysunek 12.3. Wykres funkcji daszkowej ϕ_{kl}.

na trójkącie τ. Powiedzmy, że wzory wyznaczymy na tym trójkącie, który jest zaznaczony na rysunku 12.1 - przyjmujemy, że $x^{(kl)} = (k * h, l * h)$ jest wierzchołkiem przy kącie prostym:

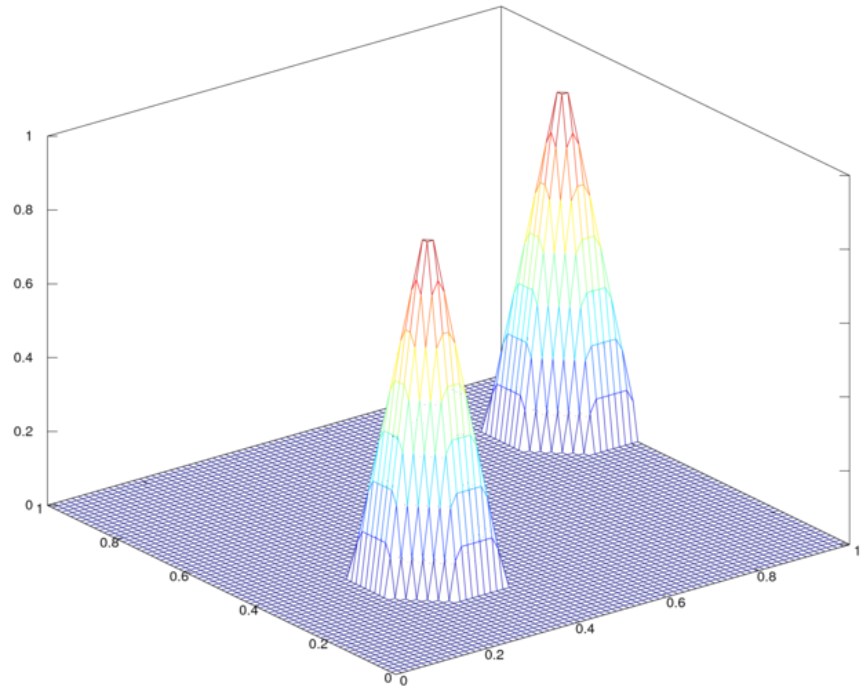
$$\begin{aligned} \phi_{k,l}(x) &= 1 - \frac{x_1 - k * h}{h} - \frac{x_2 - l * h}{h}, \\ \phi_{k+1,l}(x) &= \frac{x_1 - k * h}{h}, \\ \phi_{k,l+1}(x) &= \frac{x_2 - l * h}{h}. \end{aligned} \tag{12.2}$$

dla $x = (x_1, x_2)$. Wykresy takiej funkcji dla tej regularnej triangulacji kwadratu lub kilku funkcji możemy obejrzeć na rysunkach 12.3 i 12.4. Na rysunku 12.5 widzimy przykładową funkcję z przestrzeni V^h. Również jako zadanie pozostawimy wykazanie, że

$$\{\phi_{k,l}\}_{k,l=1,\ldots,N-1}$$

tworzą bazę V^h, i że

$$u = \sum_{k,l} u(x^{(kl)})\phi_{k,l}.$$



Rysunek 12.4. Wykresy dwóch funkcji daszkowych.

Wprowadzamy zadanie dyskretne: chcemy znaleźć $u_h \in V^h$ takie, że

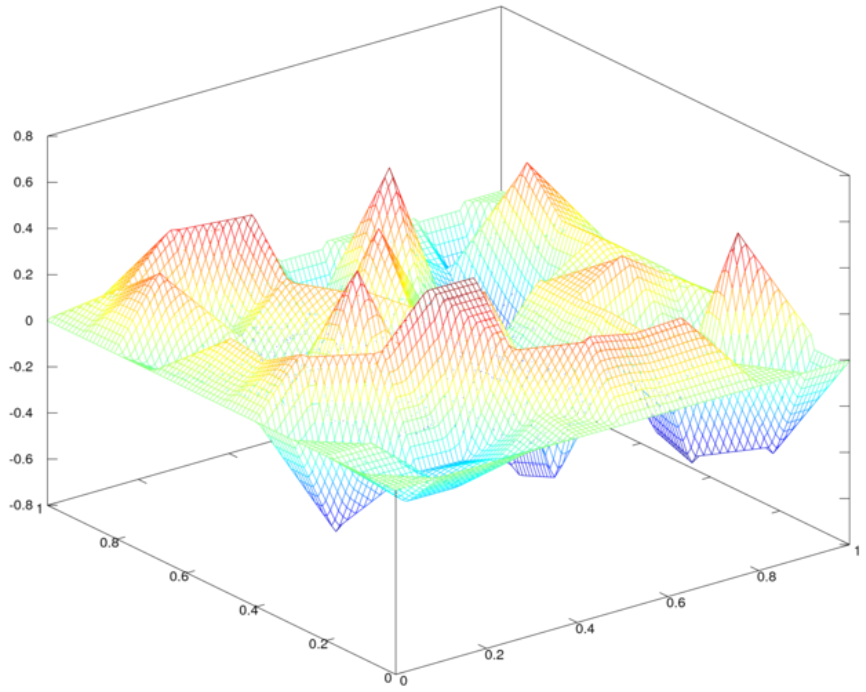
$$\int_\Omega \nabla u_h \nabla v_h + c u_h \, v_h \, dx = \int_\Omega f v_h \, dx \qquad \forall v_h \in V^h. \tag{12.3}$$

Można pokazać, że to zadanie ma jednoznaczne rozwiązanie, i że jeśli $u_* \in C^2([0,1]^2)$ to:

$$\begin{aligned} \|\nabla(u_h - u_*)\|_{L^2(\Omega)} &= O(h) \\ \|u_h - u_*\|_{L^2\Omega)} &= O(h^2). \end{aligned} \tag{12.4}$$

Zatem w normie L^2 widzimy oszacowanie zbieżności rzędu dwa analogiczne jak dla dyskretyzacji tego samego zadania przy pomocy metody różnic dzielonych i dyskretnej normy L^2_h, ale przy dużo słabszych założeniach. Tam musieliśmy założyć, że funkcja jest klasy C^4.

Można też zauważyć, że w przypadku bardziej skomplikowanego geometrycznie obszaru (wielokąta) możemy skonstruować analogiczną dyskretyzację wprowadzając triangulację złożoną z trójkątów, a w przypadku różnic dzielonych, jeśli obszar nie jest prostokątem, pojawiają się kłopoty z postawieniem warunku brzegowego.



Rysunek 12.5. Wykres przykładowej funkcji z przestrzeni elementu skończonego.

12.1.3. Element kwadratowy i kubiczny

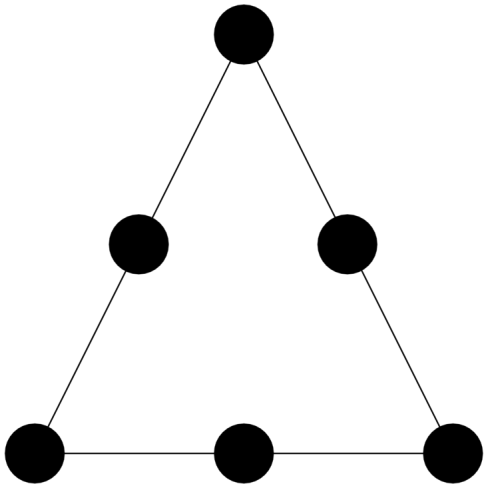
W tym rozdziale przedstawimy dwa typy przestrzeni elementu skończonego wyższych rzędów - element kwadratowy i kubiczny. Rozpatrzmy triangulację kwadratu jednostkowego $T_h([0,1]^2)$ jak w Rozdziale 12.1.1. Wtedy definiujemy:

$$V_p^h = \{u \in C(\overline{\Omega}) : u_{|\tau} \in P_p(\tau) \quad \forall \tau \in T_h(\Omega), \quad u(s) = 0 \quad \forall s \in \partial\Omega\}.$$

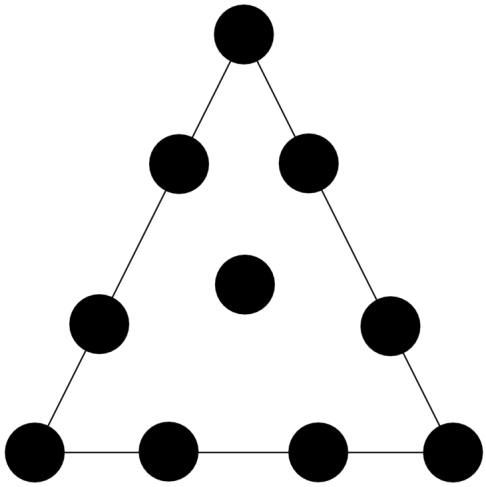
dla $p = 1, 2, 3, \ldots$. Tutaj $P_p(\tau) = \text{span}(x_1^k x_2^l)_{0 \leqslant k+l \leqslant p}$ to przestrzeń wielomianów na τ stopnia nie większego od p.

Przypadek V_2^h określamy jako przestrzeń elementu skończonego kwadratowego, a V_3^h kubicznego.

Dla $p = 2$ widzimy, że wielomian kwadratowy na trójkącie jest określony jednoznacznie poprzez swoje wartości w trzech wierzchołkach i trzech środkach krawędzi. Wszystkie te punkty wewnątrz Ω określamy jako punkty nodalne V_2^h. Z każdym takim punktem x wiążemy funkcję, która jest równa jeden w tym punkcie, a zero we wszystkich pozostałych wierzchołkach i punktach środkowych krawędzi, por. rysunek 12.6. Z kolei dla $p = 3$ wielomian kubiczny na trójkącie jest jednoznacznie określony poprzez swoje wartości w wierzchołkach, w środku ciężkości oraz w dwu punktach wewnątrz każdej krawędzi dzielących ją na trzy równe odcinki, por. rysunek 12.7.



Rysunek 12.6. Wierzchołki i punkty środkowe krawędzi trójkąta τ -punkty swobody wielomianu kwadratowego na tym trójkącie.



Rysunek 12.7. Punkty swobody wielomianu kubicznego na trójkącie τ.

W przypadku przestrzeni V_p^h możemy wprowadzić analogiczne bazy nodalne złożone z funkcji, które przyjmują wartość jeden w ustalonym punkcie nodalnym, a zero - w pozostałych. Wzory takich funkcji są coraz bardziej skomplikowane wraz ze wzrostem p. Dla $p = 2$ wypiszemy wzory na obcięcie funkcji nodalnej na ustalonym trójkącie τ. Niech punkty $x^{(k)}$ $k = 0, 1, 2$ będą trzema wierzchołkami tego trójkąta τ i niech punkt $m^{(k)}$ $k = 0, 1, 2$ będzie środkiem odcinka między $x^{(k)}$ a $x^{(k+1)}$. Przy czym utożsamiamy 0 z 3. Niech q_k będzie funkcją liniową taką, że $q_k(x^{(j)}) = 0$ dla $j \neq k$ i $q_k(x^{(k)}) = 1$. Wtedy funkcja nodalna związana z wierzchołkiem elementu $x^{(k)}$ dla elementu kwadratowego wynosi na trójkącie τ:

$$\phi_{x^{(k)}} = q_k(q_k - q_{k-1} - q_{k+1}) \qquad k = 0, 1, 2 \; mod \; 3, \tag{12.5}$$

a funkcja nodalna związana z punktem środkowym krawędzi elementu na τ:

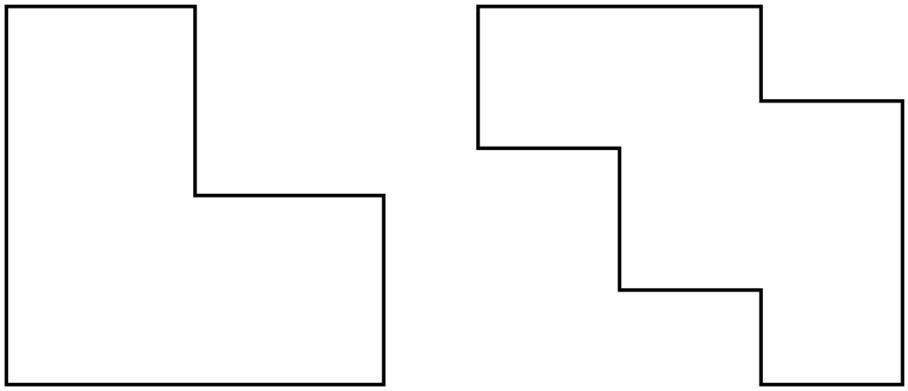
$$\phi_{m^{(k)}} = 4 * q_k * q_{k+1} \qquad k = 0, 1, 2 \; mod \; 3. \tag{12.6}$$

Jako zadanie pozostawiamy sprawdzenie tych wzorów i znalezienie analogicznych dla funkcji bazowych elementu kubicznego. Zastosowanie elementów: kwadratowego ($p = 2$), kubicznego

($p = 3$) czy nawet dla większych p jest korzystne, o ile u_* - czyli rozwiązanie zadania różniczkowego (12.1) jest bardziej regularne, tzn. należy do $C^{s+1}(\overline{\Omega})$ dla s większych od jeden do $s = p$. Tzn. wtedy można wykazać, że

$$\|\nabla(u_h - u_*)\|_{L^2(\Omega)} = O(h^s).$$

12.1.4. Metoda elementu skończonego z podziałem obszaru na prostokąty



Rysunek 12.8. Przykład obszarów będących sumą prostokątów.

W przypadku, gdy nasz obszar jest prostokątem lub kwadratem, możemy wprowadzić przestrzeń elementu skończonego: tzw. elementów skończonych prostokątnych. Taki element możemy w ogólności zastosować, jeśli obszar jest sumą prostokątów o brzegach będących sumą odcinków równoległych do osi współrzędnych; np. tzw. L-obszarem, por. rysunek 12.8.

Opiszemy tę metodę dla naszego modelowego zadania postawionego na kwadracie. Dla $\Omega = (0,1)^2$ wprowadźmy triangulacje złożoną z kwadratów, por. rysunek 12.9:

$$T_h(\Omega) = \{\tau_{kl}\}_{k,l=0,\ldots,N-1}$$

dla $h = 1/N$ i dla kwadratów:

$$\tau_{kl} = (k * h, (k+1) * h) \times (l * h, (l+1) * h),$$

Wtedy przestrzeń dyskretną definiujemy jako przestrzeń funkcji ciągłych biliniowych na kwadratach:

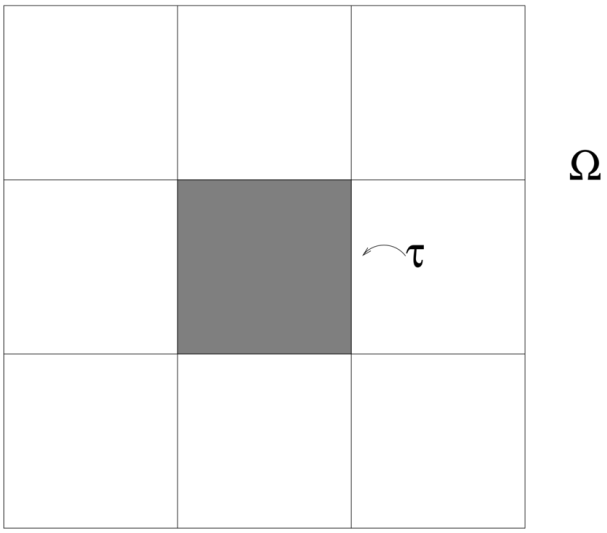
$$V_{bi}^h = \{u \in C(\overline{\Omega}) : u_{|\tau} \in Q_1(\tau) \quad \forall \tau \in T_h(\Omega), \quad u(s) = 0 \quad \forall s \in \partial\Omega\}.$$

gdzie

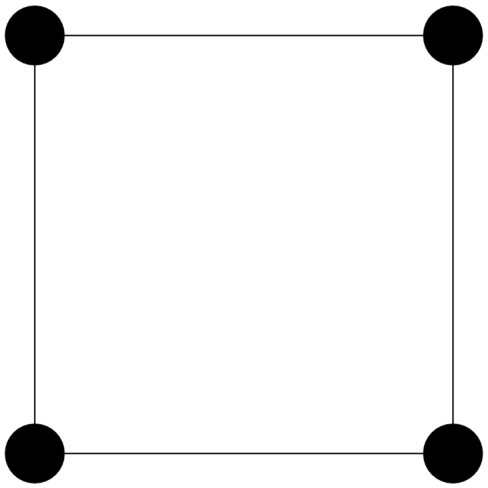
$$Q_1(\tau_{kl}) = P_1((k * h, (k+1) * h)) \otimes P_1((l * h, (l+1) * h)) = \mathrm{Span}(1, x_1, x_2, x_1\, x_2)$$

przestrzeń wielomianów biliniowych określonych na τ_{kl}, czyli liniowych ze względu na każdą zmienną z osobna.

Wielomian biliniowy na prostokącie jest zdefiniowany poprzez wartości w wierzchołkach tego prostokąta. Nazywamy je punktami swobody tego elementu skończonego, por. rysunek 12.10. Analogicznie do przypadku elementów trójkątnych możemy tu wprowadzić funkcje nodalne: definiując $\phi_{k,l} \in V_{bi}^h$ dla $k, l \in \{1, \ldots, N-1\}$ jako funkcję, która spełnia $\phi_k(x^{(kl)}) = 1$ i



Rysunek 12.9. Przykład triangulacji kwadratu jednostkowego złożonej z kwadratów.



Rysunek 12.10. Wierzchołki kwadratu τ -punkty swobody wielomianu biliniowego na tym kwadracie.

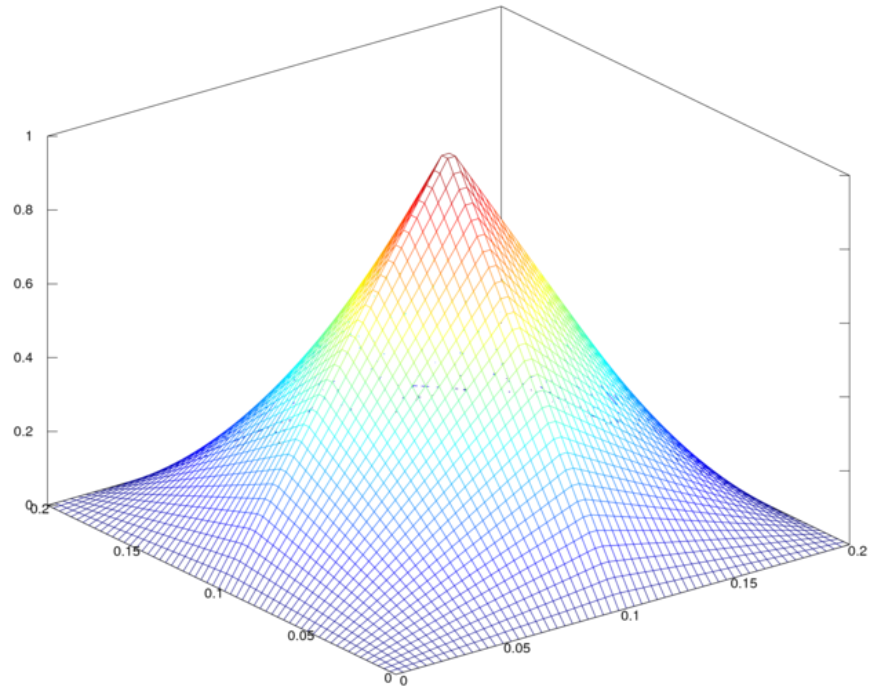
$\phi_k(x^{(ij)}) = 0$ dla $x^{(kl)} \neq x^{(ij)}$ dla $x^{(kl)} = (x^{(k)}, x^{(l)})$. Podanie wzoru na te funkcje jest nietrudnym zadaniem z uwagi na prostą geometrię elementu. Pozostawiamy to jako zadanie, por. rysunek 12.11 Następnie definiujemy zadanie dyskretne: chcemy znaleźć $u_{hb} \in V_{bi}^h$ takie, że

$$\int_\Omega \nabla u_{hb} \nabla v_h + c u_h \, v_h \, dx = \int_\Omega f v_h \, dx \qquad \forall v_h \in V_{bi}^h. \tag{12.7}$$

W tym przypadku możemy analogicznie wykazać, że zachodzi oszacowanie błędu takie, jak w przypadku elementu liniowego na trójkącie. Tzn. można wykazać, że to zadanie ma jednoznaczne rozwiązanie, i że jeśli $u_* \in C^2([0,1]^2)$, to

$$\begin{aligned} \|\nabla(u_{hb} - u_*)\|_{L^2(\Omega)} &= O(h) \\ \|u_{hb} - u_*\|_{L^2\Omega)} &= O(h^2). \end{aligned} \tag{12.8}$$

W przypadku zadania na kwadracie jednostkowym element biliniowy wydaje się bardziej naturalny, ale przestrzeń elementu skończonego jest bardziej skomplikowana, ponieważ na każdym kwadracie triangulacji lokalna przestrzeń MESu zawiera wielomiany wyższego stopnia.



Rysunek 12.11. Wykres funkcji bazowej przestrzeni MES biliniowej na kwadracie.

Czy możemy uzyskać lepsze oszacowania błędu? Nie - jeśli chodzi o rząd zbieżności, tzn. dla elementu biliniowego otrzymamy co najwyżej $O(h)$ w normie $\|\nabla \cdot\|_{L^2(\Omega)}$, ale oszacowanie błędu jest ostrzejsze, por. rozdział 4.6 w [4].

12.2. Niejednorodny warunek brzegowy

Można się zastanowić co się dzieje, jeśli w (7.8) w warunku brzegowym Dirichleta wartość prawej strony jest różna od zera, tzn. $u_* = g$ na brzegu. Tak jak poprzednio, otrzymujemy nowe słabe sformułowanie: chcemy znaleźć u_* takie, że $u_* = g$ na brzegu Ω spełniające (12.1). Jeśli założymy, że znamy funkcję $\tilde{g}$ określoną na Ω taką, że $\tilde{g} = g$ na brzegu, to definiując $\hat{u}_* = u_* - \tilde{g}$ otrzymujemy:

$$a(\hat{u}_*, v) = (f, v) - a(\tilde{g}, v) = F(v) \qquad \forall v \in V,$$

dla V jak w (12.1) i $\hat{u}^* \in V$, czyli (12.1) ale z prawą stroną zależną dodatkowo od $\tilde{g}$.

Następnie możemy wprowadzić zadanie dyskretne tak, jak dla zerowych warunków brzegowych.

12.3. Zadania

Ćwiczenie 12.1. Udowodnij, że funkcja bazowa przestrzeni elementu liniowego na trójkącie spełnia wzory (12.3).

Ćwiczenie 12.2. Definiujemy rozwiązanie (12.1) jako

$$u_h = \sum_{k,l=0}^{N-2} u_{k+(N-1)*l}\ \phi_{k+1,l+1}$$

z $u(x^{(k+1,l+1)}) = u_{k+(N-1)*l}$. Wstawiając u_h w tej postaci do (12.1) otrzymujemy układ równań liniowych na wektor współrzędnych $\vec{u} = (u_{k+(N-1)*l})_{k,l=0}^{N-2}$. Policz różne od zera elementy macierzy tego układu A_h i elementy wektora prawej strony f_h. Czy elementy na diagonali tej macierzy zależą od h? Pokaż, że dla równomiernego podziału, tzn. $x^{(k)} = k * h$, macierz ta jest równa macierzy dyskretyzacji metodą różnic skończonych dla tego samego zadania, pomnożonej przez parametr h^2, tzn. macierzą układu równań liniowych dla ćwiczenia 7.5 (dla $c = 0$). Czy oba układy po przeskalowaniu przez h są wtedy identyczne?

Rozwiązanie. Układy równań liniowych nie są identyczne, ponieważ prawe strony są różne, tak jak w przypadku jednowymiarowym.

Ćwiczenie 12.3. Pokaż, że macierz A_h z ćwiczenia 12.2 jest zawsze pasmowa (wyznacz wielkość pasma, wyznacz czy zależy ono od N, czyli odpowiednio od h), symetryczna i dodatnio określona.

Ćwiczenie 12.4. Zakładając, że posiadamy procedurę z metodą iteracyjną, która rozwiązuje układ równań z ćwiczenia 12.2 w $I(N)$ iteracji, i dla której w każdej iteracji wykonujemy jedno mnożenie przez macierz A_h oraz $10 * N^2$ operacji algebraicznych określ, ile musi wynosić $I(N)$, aby metoda ta była tańsza od odpowiedniej taśmowej wersji eliminacji Gaussa dla macierzy symetrycznej dodatnio określonej.

Ćwiczenie 12.5. Pokaż, że dla dowolnej funkcji przestrzeni liniowej elementu skończonego $u_h \in V^h$, por. rozdział 12.1.2, zachodzi tzw. nierówność odwrotna:

$$\|\nabla u_h\|_{L^2(\Omega)} \leqslant C\, h^{-1} \|u_h\|_{L^2(\Omega)}.$$

Tutaj stała C jest niezależna od parametru h i u_h.

Wskazówka. Wystarczy pokazać to oszacowanie na dowolnym elemencie triangulacji $\tau \in T_h$ i wykorzystać fakt, że $\|\nabla u_h\|_{L^2(\tau)} = \|\nabla(u_h - c)\|_{L^2(\tau)}$ dla dowolnej stałej c.

Ćwiczenie 12.6. (laboratoryjne) Napisz w octave odpowiednią wersję eliminacji Gaussa dla macierzy pasmowej symetrycznej dodatnio określonej. Zastosuj ją do układu równań liniowych z ćwiczenia 12.2 dla N różnej wielkości. Porównaj czas w octavie z czasem dla standardowej metody rozwiązywania równań liniowych octave'a zarówno, gdy macierz układu jest w formacie pełnym, jak i formacie rzadkim (można użyć funkcje octave: **sparse**(), **tic**() i **toc**() i operator octave: \).

Ćwiczenie 12.7. Udowodnij, że funkcja z P_2 jest jednoznacznie wyznaczona przez określenie jej wartości w wierzchołkach i środkach krawędzi trójkąta (jak opisano w rozdziale 12.1.3, por. rysunek 12.6).

Ćwiczenie 12.8. Udowodnij, że funkcja z P_3 jest jednoznacznie wyznaczona przez określenie jej wartości jak opisano w rozdziale 12.1.3, por. rysunek 12.7.

Ćwiczenie 12.9. Dla przestrzeni V_2^h wyprowadź układ równań liniowych:

$$A_{h,2}\vec{u} = \vec{f}$$

taki, że jego rozwiązaniem są współczynniki rozwiązania przybliżonego $u_{h,p}$ w bazie nodalnej tej przestrzeni elementu skończonego, por. rozdział 12.1.3. Czy macierz tego układu jest pasmowa? Jeśli tak, to znajdź wielkość pasma. Czy jest symetryczna i dodatnio określona?

Ćwiczenie 12.10. Udowodnij wzory na bazowe funkcje elementu kwadratowego na trójkącie (12.5) i (12.6).

Ćwiczenie 12.11. Wyznacz wzory na bazowe funkcje nodalne dla elementu kubicznego, analogiczne do wzorów (12.5) i (12.6) dla funkcji bazowych elementu kwadratowego, tzn. wzory na te funkcje w zależności od funkcji q_k.

Ćwiczenie 12.12. Wyznacz wzory na współczynniki bazowych funkcji nodalnych dla elementu biliniowego na kwadracie $x^{(kl)} + (0,h)^2$ w bazie $(1, (x_1 - x^{(k)}), (x_2 - x^{(l)}), (x_1 - x^{(k)}) * (x_2 - x^{(l)}))$. Oblicz elementy macierzy układu równań liniowych dla dyskretyzacji metodą elementu skończonego biliniowego dla zadania dyskretnego (12.7) w tej bazie nodalnej.

Ćwiczenie 12.13. Dla przestrzeni V_{bi}^h, wyprowadź układ równań liniowych:

$$A_{h,bi}\vec{u} = \vec{f}$$

taki, że jego rozwiązaniem są współczynniki rozwiązania przybliżonego u_{hb} w bazie nodalnej związanej z wierzchołkami elementów kwadratowych. Czy macierz tego układu jest pasmowa? Jeśli tak, to znajdź wielkość pasma. Czy jest symetryczna i dodatnio określona?

13. Metoda elementu skończonego - teoria

W tym wykładzie przedstawimy ogólną teorię konstrukcji i analizy zbieżności elementu skończonego (MESu) dla równań liniowych.

13.1. Istnienie rozwiązania

Załóżmy, że V jest rzeczywistą przestrzenią Hilberta, tzn. rzeczywistą przestrzenią liniową z iloczynem skalarnym $(\cdot,\cdot)$ i normą $\|\cdot\|_V = (\cdot,\cdot)^{1/2}$, która jest zupełna. Przez V^* oznaczamy przestrzeń dualną (sprzężoną) do V, por. np. [7].

Rozpatrzmy wariacyjny problem znalezienia $u^* \in V$ takiego, że

$$a(u^*, v) = f(v) \qquad \forall v \in V, \tag{13.1}$$

gdzie $f \in V^*$, $a(u,v)$ jest formą dwuliniową, która jest ograniczona, tzn. istnieje stała $M \geqslant 0$ taka, że

$$a(u,v) \leqslant M\|u\|_V\|v\|_V \qquad \forall u, v \in V,$$

oraz jest V-eliptyczna co oznacza, że dla pewnego $\alpha > 0$ zachodzi:

$$a(u,u) \geqslant \alpha\|u\|_V^2 \qquad \forall u \in V.$$

Przy powyższych założeniach zachodzi znane twierdzenie analizy funkcjonalnej:

Twierdzenie 13.1 (Lax-Milgram)**.** *Rozpatrzmy formę dwuliniową $a(u,v) : V \times V \to \mathbb{R}$, która jest ograniczona i V-eliptyczna, a $f \in V^*$. Wtedy zadanie (13.1) ma jednoznaczne rozwiązanie i*

$$\|u^*\|_V \leqslant \frac{\|f\|_{V^*}}{\alpha}. \tag{13.2}$$

Dowód. Istnienie rozwiązania wynika z lematu Riesza. Szczegóły dowodu można znaleźć np. w [7] lub [6]. Oszacowanie (13.2) dla u^* otrzymujemy wstawiając u^* za v w (13.1) i korzystając z V-eliptyczności formy i definicji normy dualnej funkcjonału liniowego:

$$\alpha\|u^*\|_V^2 \leqslant a(u^*,u^*) = f(u^*) \leqslant \|f\|_{V^*}\|u^*\|_V.$$

Jeśli u_1 i u_2 są rozwiązaniami zadania wariacyjnego, to $w = u_1 - u_2$ spełnia (13.1) dla prawej strony równej zero. Z tego i z (13.2) wynika, że $\|u_1 - u_2\|_V = 0$, co oznacza, że rozwiązanie jest wyznaczone jednoznacznie. □

13.2. Metoda Galerkina

Załóżmy, że $\{V^n\}_n$ to rodzina podprzestrzeni skończenie wymiarowych V o wymiarze n.

Definiujemy zadanie dyskretne aproksymujące (13.1): chcemy znaleźć $u_n \in V^n$ takie, że

$$a(u_n^*, v_n) = f(v_n) \qquad \forall v_n \in V^n. \tag{13.3}$$

Forma $a(u,v)$ jest ograniczona na V^n z normą przestrzeni V i jest również V^n-eliptyczna. Zatem z twierdzenia 13.1 wynika istnienie jednoznacznego rozwiązania zadania dyskretnego, które spełnia:

$$\|u_n^*\|_V \leqslant \frac{\|f\|_{V^*}}{\alpha}, \tag{13.4}$$

Rozwiązania dyskretne są wspólnie ograniczone niezależnie od wymiaru n, co określamy jako stabilność rozwiązań rodziny zadań dyskretnych.

Proszę zauważyć, że ponieważ przestrzeń V^n jest skończenie wymiarowa, więc - z definicji - ma bazę o skończonej ilości elementów $n < \infty$, tzn. $V^n = \{\phi_j\}_{j=1}^n$ i, aby znaleźć współczynniki rozwiązania (13.3) w tej bazie, należy rozwiązać układ równań liniowych

$$A_n \vec{u} = \vec{f},$$

gdzie $(A_n)_{kl} = a(\phi_k, \phi_l)$ i $\vec{f} = (f(\phi_k))_k$ dla $k, l = 1, \ldots, n$. Jeśli forma $a(u,v)$ jest symetryczna, to A_n jest macierzą symetryczną i dodatnio określoną. Oczywiście najlepiej byłoby dobrać taką bazę, żeby macierz A_n była np. pasmowa, albo ogólniej - o małej ilości elementów różnych od zera. Pojawia się pytanie: jak taką bazę wyznaczyć?

13.3. Abstrakcyjne oszacowanie błędu

Tutaj pokażemy związek między błędem $u^* - u_n^*$, a błędem aproksymacji przestrzeni V przez rodzinę przestrzeni V^n.

Zachodzi ważne twierdzenie - zwyczajowo zwane lematem Céa:

Twierdzenie 13.2 (lemat Céa)**.** *Niech forma $a(u,v)$ określona na przestrzeni Hilberta V będzie ograniczona i V-eliptyczna, $V^n \subset V$ podprzestrzeń V. Wtedy*

$$\|u^* - u_n^*\|_V \leqslant \frac{M}{\alpha} \inf_{v_n \in V^n} \|u^* - v_n\|_V,$$

gdzie u^* - to rozwiązanie (13.1), a u_n^* - to rozwiązanie (13.3).

Dowód. Z (13.1) wynika, że:

$$a(u^*, v_n) = f(v_n) \qquad \forall v_n \in V^n.$$

Odejmując to równanie od (13.3) otrzymujemy:

$$a(u^* - u_n^*, v_n) = 0 \qquad \forall v_n \in V^n.$$

A dalej

$$\begin{aligned} \alpha\|u^* - u_n^*\|_V^2 &\leqslant a(u^* - u_n^*, u^* - u_n^*) = a(u^* - u_n^*, u^*) - a(u^* - u_n^*, u_n^*) \\ &= a(u^* - u_n^*, u^*) - a(u^* - u_n^*, v_n) = a(u^* - u_n^*, u^* - v_n) \\ &\leqslant M\|u^* - u_n^*\|_V \|u^* - v_n\|_V. \end{aligned}$$

Dzieląc przez $\|u^* - u_n^*\|_V$ otrzymujemy tezę twierdzenia. □

Z lematu wynika, że aby oszacować błąd $u^* - u_n^*$ wystarczy oszacować błąd aproksymacji u^* przez podprzestrzeń dyskretną V^n.

Wniosek 13.1. *Przy założeniach lematu Céa, jeśli rodzina podprzestrzeni $\{V^n\}$ przestrzeni Hilberta V jest taka, że:*

$$\forall u \in V \quad \operatorname{dist}(u, V^n) = \inf_{v \in V^n} \|u - v\|_V \to 0 \qquad n \to \infty,$$

to

$$\|u_n^* - u^*\|_V \to 0 \qquad n \to \infty.$$

13.4. Przestrzenie Sobolewa

Niech $\Omega \subset \mathbb{R}^d$ będzie otwartym obszarem. Wtedy definiujemy półnormę i normę $H^1(\Omega)$ jako:

$$|u|_{H^1(\Omega)} = \sqrt{\int_\Omega |\nabla u|^2 \, dx}, \qquad \|u\|_{H^1(\Omega)} = \sqrt{\|u\|^2_{L^2(\Omega)} + |u|^2_{H^1(\Omega)}}.$$

Dodatkowo będziemy też oznaczać $H^0(\Omega) := L^2(\Omega)$ i $|u|_{H^0(\Omega)} = \|u\|_{H^0(\Omega)} = \|u\|_{L^2(\Omega)}$.

Definicja 13.1. Niech $H_0^1(\Omega) \subset L^2(\Omega)$ będzie domknięciem w normie H^1 przestrzeni $C_0^\infty(\Omega)$, gdzie $C_0^\infty(\Omega)$ jest podprzestrzenią $C^\infty(\Omega)$ złożoną z funkcji o zwartym nośniku w Ω.

Jest to przestrzeń zupełna (ang. *complete space*) z iloczynem skalarnym H^1: $(u, v)_{H^1(\Omega)} = \int_\Omega u\, v + \nabla u \, \nabla v \; dx$. Można pokazać, że jest to ośrodkowa przestrzeń Hilberta (ang. *separable Hilbert space*), tzn. posiada przeliczalną bazę ortonormalną (ang. *countable orthonormal basis*).

Pojawia się pytanie; czy jeśli funkcja $u \in H_0^1(\Omega) \cap C(\overline{\Omega})$, to $u_{|\partial\Omega} = 0$. Okazuje się, że tak jest, co wynika z twierdzenia o śladzie, por. twierdzenie 16.2.

Z nierówności Friedrichsa (por. stwierdzenie 16.1) wynika, że półnorma H^1 w przestrzeni $H_0^1(\Omega)$ jest normą równoważną z normą H^1.

Dodatkowo zdefiniujemy półnormę H^2:

$$|u|_{H^2(\Omega)} = \sqrt{\int_\Omega \sum_{k,l=1}^{d} \left| \frac{\partial^2 u}{\partial x_k \partial x_l} \right|^2 dx}.$$

poprawnie zdefiniowaną dla funkcji gładkich i przestrzeń $H_0^1(\Omega) \cap H^2(\Omega)$ złożoną z tych funkcji w $H_0^1(\Omega)$, dla których jej drugie pochodne dystrybucyjne są w $L^2(\Omega)$. Przestrzeń ta zawiera wszystkie funkcje klasy $C^2(\overline{\Omega})$ zerujące się na brzegu Ω.

13.5. Zadanie eliptyczne drugiego rzędu z zerowymi warunkami na brzegu

Rozpatrzmy ogólne zadanie eliptyczne w słabym sformułowaniu: chcemy znaleźć $u^* \in H_0^1(\Omega)$ takie, że

$$a(u^*, v) = f(v) \qquad \forall v \in H_0^1(\Omega), \tag{13.5}$$

gdzie $f(v) = \int_\Omega \hat{f} v \, dx$ dla danej funkcji $\hat{f} \in L^2(\Omega)$ oraz

$$a(u, v) = \int_\Omega \sum_{i,j=1}^{d} a_{ij}(x) \frac{\partial u}{\partial x_i} \frac{\partial v}{\partial x_j} + c(x) u\, v \, dx,$$

Tutaj funkcje $a_{ij}, c \in L^\infty(\Omega)$, tzn. są ograniczone, oraz istnieje stała $\hat{\alpha} > 0$ taka, że:

$$\begin{aligned} \sum_{i,j=1}^{d} a_{ij}(x)\xi_i\xi_j &\geqslant \hat{\alpha}\sum_{i=1}^{d}\xi_i^2, \qquad \forall \xi \in \mathbb{R}^d, \quad \forall x \in \Omega \\ c(x) &\geqslant 0 \qquad \forall x \in \Omega. \end{aligned}$$

Jeśli dodatkowo $a_{ij} = a_{ji}$ na Ω to mówimy, że zadanie jest samosprzężone. Można wykazać, że istnieją stałe dodatnie M, α takie, że:

$$\begin{aligned} |a(u,v)| &\leqslant M\|u\|_{H^1(\Omega)}\|v\|_{H^1(\Omega)} \qquad \forall u, v \in H_0^1(\Omega), \\ a(u,u) &\geqslant \alpha\|u\|_{H^1(\Omega)} \qquad \forall u \in H_0^1(\Omega), \end{aligned} \tag{13.6}$$

czyli forma dwuliniowa $a(\cdot,\cdot)$ jest ograniczona w $H_0^1(\Omega)$ i $H_0^1(\Omega)$-eliptyczna.

Jako wniosek z twierdzenia 13.1 otrzymujemy:

Stwierdzenie 13.1. *Zadanie (13.5) ma jednoznaczne rozwiązanie.*

Jeśli zadanie jest samosprzężone to:

$$a(u,v) = a(v,u)$$

i forma $a(\cdot,\cdot)$ jest iloczynem skalarnym w $H_0^1(\Omega)$.

Można pokazać, że jeśli istnieje rozwiązanie u^* zadania (13.5), które dodatkowo jest klasy $C^2(\Omega)$, i jeśli funkcje $a_{ij} \in C^1(\Omega)$, to

$$-\sum_{k,l=1}^{n}\frac{\partial}{\partial x_k}\left(a_{kl}(x)\frac{\partial u}{\partial x_l}(x)\right) + c(x)u(x) = f(x).$$

13.6. Ciągła metoda elementu skończonego dla zadań eliptycznych drugiego rzędu

W tym rozdziale przedstawimy ogólne zasady konstrukcji ciągłej metody elementu skończonego. Ciągłość oznacza, że przestrzenie elementu skończonego będą zawierały wyłącznie funkcje ciągłe z przestrzeni wyjściowej V.

Będziemy zajmowali się konstrukcją przestrzeni wyłącznie dla zagadnień różniczkowych zadanych na ograniczonym obszarze $\Omega \subset \mathbb{R}^d$ dla $d = 1, 2, 3$.

13.6.1. Triangulacje

Będziemy zakładali, że Ω jest odcinkiem dla $d = 1$, wielokątem dla $d = 2$, czy wielościanem dla $d = 3$.

Wprowadzamy w Ω rodzinę podziałów $T_h(\Omega) = \{\tau\}$ dla $\tau \subset \Omega$ na odpowiednio: odcinki dla $d = 1$, trójkąty lub prostokąty dla $d = 2$, czworościany lub prostopadłościany dla $d = 3$ - przy czym typ elementu zawsze jest ustalony. Formalnie wprowadzamy następującą definicję triangulacji, por. rozdział 12.1.1:

Rozpatrzmy obszar $\Omega \subset \mathbb{R}^d$, $d = 1, 2, 3$ będący odcinkiem, wielokątem (ang. *polygon*), lub wielościanem (ang. *polyhedron*), i niech $T(\Omega) = \{\tau_1, \ldots, \tau_n\}$ będzie podziałem Ω, tzn. rodziną wielościanów (ang. *polyhedrons or elements*) zazwyczaj ustalonego typu, tzn. odcinków (ang. *segments*) dla $d = 1$, trójkątów (ang. *triangles*), czworokątów (ang. *quadrilaterals*) lub prostokątów (ang. *rectangles*) dla $d = 2$, czworościanów (ang. *tetrahedrons*) lub prostopadłościanów (ang. *cuboids*) czy sześcianów (ang. *cubes*) dla $d = 3$.

Definicja 13.2. (Triangulacja obszaru)

1. Powiemy, że $T_h = T_h(\Omega)$ jest dopuszczalną triangulacją (ang. *admissible triangulation*), jeśli spełnione są następujące warunki:
   — $\bigcup_{\tau\in T(\Omega)} \overline{\tau} = \overline{\Omega}$,
   — $\partial\tau_k \cap \partial\tau_l$ jest zbiorem pustym, wspólnym wierzchołkiem, wspólną krawędzią ($d = 2, 3$), wspólną ścianą (tylko $d = 3$), jeśli $k \neq l$.
2. Dla danej triangulacji $T_h(\Omega)$ niech $h = \max_{\tau\in T(\Omega)} \operatorname{diam}(\tau)$ oznacza parametr tej triangulacji.
3. Rodzina triangulacji $\{T_h(\Omega)\}$ jest regularna ze względu na kształt (ang. *shape regular*), jeśli istnieje stała $c > 0$ taka, że każdy τ w T_h zawiera okrąg wpisany w τ o promieniu ρ_τ taki, że
$$\rho_\tau \geqslant c \operatorname{diam}(\tau).$$
4. Rodzina triangulacji $\{T_h(\Omega)\}$ jest regularna równomiernie (ang. *quasiuniform*), jeśli jest regularna ze względu na kształt i istnieje stała C taka, że każdy τ w T_h zawiera okrąg wpisany w τ o promieniu ρ_τ taki, że
$$\rho_\tau \geqslant C\, h.$$

Własność regularności ze względu na kształt i własność równomiernej regularności są niezbędne w teorii zbieżności metod elementu skończonego.

Będziemy zakładali, że dla rodziny triangulacji $\{T_h(\Omega)\}$ - czyli podziałów na wielościany (ang. *polyhedrons*) ustalonego typu - istnieje tzw. wielościan wzorcowy $\hat{\tau}$ i ustalona przestrzeń wielomianów $\hat{P}$ określonych na $\hat{\tau}$ wraz z ustalonymi różnymi punktami $\hat{x}_j \in \overline{\hat{\tau}}$ takimi, że każdy wielomian $p \in \hat{P}$ jest wyznaczony jednoznacznie przez swoje wartości w tych punktach.

Definicja 13.3. Rozpatrzmy rodzinę przestrzeni funkcji ciągłych $\{V^h\}_h$ takich, że dla triangulacji T_h i dowolnego $\tau_j \in T_h$ istnieje izomorficzne przekształcenie afiniczne $F_j : \hat{\tau} \to \tau_j$ takie, że dla dowolnej funkcji $u \in V^h$ istnieje $w \in \hat{P}$ takie, że
$$u(x) = w(F_j^{-1}x), \qquad x \in \tau_j.$$

Wtedy V^h nazywamy przestrzenią ciągłego elementu skończonego (ang. *continuous finite element space*), a rodzinę tych przestrzeni - afiniczną rodziną ciągłych przestrzeni elementu skończonego (ang. *affine family of FE spaces*). Punkty $x_k = F_j(\hat{x}_k) \in \tau$ nazywamy punktami nodalnymi na elemencie τ_j, a ich zbiór oznaczamy $N_h(\tau_j)$, a przestrzeń wielomianów $P(\tau_j) = \{u_{|\tau_j} : u \in V^h\} = \{u : \exists\, w \in \hat{P},\ u(x) = w(F_j^{-1}x),\ x \in \tau_j\}$ jest lokalną przestrzenią wielomianów na τ_j.

Zbiór punktów nodalnych (ang. *nodal points*)

Dodatkowo będziemy zakładali, że dla danej przestrzeni ciągłej elementu skończonego V^h zbudowanej na triangulacji T_h obszaru Ω istnieje $N_h \subset \bigcup_{\tau\in T_h} N_h(\tau)$ - podzbiór zbioru wszystkich punktów nodalnych wielościanów z triangulacji T_h taki, że wartości funkcji z V^h, zwane wartościami nodalnymi tej funkcji w tym zbiorze, jednoznacznie tę funkcję definiują.

Wprowadzamy definicję bazy nodalnej związanej z punktami nodalnymi:

Definicja 13.4. Bazą nodalną (ang. *nodal basis*) związaną ze zbiorem punktów nodalnych N_h (ang. *nodal points*) nazywamy układ funkcji $\{\phi_x\}_{x\in N_h}$ w V^h taki, że
$$\phi_x(y) = \begin{cases} 1 & y = x \\ 0 & y \neq x \end{cases} \qquad \forall y \in N_h.$$

.

Ten układ jest bazą w V^h i widzimy, że:

$$u = \sum_{x \in N_h} u(x)\phi_x \qquad \forall u \in V^h. \tag{13.7}$$

Wprowadzamy też pojęcie operatora interpolacji nodalnej:

Definicja 13.5. Rozpatrzmy V^h ciągłą przestrzeń elementu skończonego, zbudowaną na triangulacji T_h, oraz niech N_h będzie zbiorem punktów nodalnych dla tej przestrzeni. Wtedy operatorem interpolacji nodalnej (ang. *nodal interpolant*) dla V^h nazwiemy operator: $\pi_h : C(\overline{\Omega}) \to V^h$ zdefiniowany jako

$$\pi_h u = \sum_{x \in N_h} u(x)\phi_x.$$

Nietrudno zauważyć, że:

Stwierdzenie 13.2. *Operator interpolacji π_h jest rzutem na V^h, tzn.*

$$\pi_h C(\overline{\Omega}) = V^h, \qquad \pi_h v_h = v_h \quad \forall v_h \in V^h.$$

13.6.2. Warunek ciągłości, a przestrzeń Sobolewa H_0^1

Następne twierdzenie podaje nam warunek dostateczny na to, by przestrzeń zawierająca funkcje, które na podzbiorach są odpowiednio gładkie była zawarta w $H_0^1(\Omega)$.

Twierdzenie 13.3. *Niech T_h będzie triangulacją obszaru Ω. Niech $u \in C(\overline{\Omega})$ będzie taka, że $u_{|\partial\Omega} = 0$ i $u_{|\tau} \in C^1(\tau)$ dla dowolnego $\tau \in T_h$. Wtedy $u \in H_0^1(\Omega)$.*

Dowód można znaleźć np. w [7]. Wynika z niego, że:

Wniosek 13.2. *Jeśli wszystkie funkcje z ciągłej przestrzeni elementu skończonego V^h na obszarze Ω (por. definicja 13.3) przyjmują zerowe wartości na brzegu Ω, to V^h jest podprzestrzenią $H_0^1(\Omega)$.*

13.6.3. Aproksymacyjne własności ciągłych przestrzeni elementu skończonego w H_0^1

Zachodzi następujące twierdzenie o aproksymacji dla operatora interpolacji nodalnej:

Twierdzenie 13.4. *Rozpatrzmy $\{V^h\}_h$ afiniczną rodzinę ciągłych przestrzeni elementu skończonego zbudowanych na dopuszczalnej rodzinie triangulacji regularnych co do kształtu taką, że $P_1 \subset \hat{P}$, oraz $u \in H_0^1(\Omega) \cap H^2(\Omega)$. Wtedy dla operatora interpolacji nodalnej w przestrzeni V^h zachodzi:*

$$\|u - \pi_h u\|_{L^2(\Omega)} + h|u - \pi_h u|_{H^1(\Omega)} \leqslant Ch^2|u|_{H^2(\Omega)}.$$

Dowód. Rozpatrzmy funkcję ciągła u na elemencie $\tau \in T_h$ i $\pi_{h,\tau_j} u$ taką funkcję w $P(\tau_j)$, że

$$\pi_{h,\tau_j} u(x) = u(x) \qquad x \in N_h(\tau).$$

Wtedy

$$\pi_{h,\tau_j} u(x) = \pi_h u(x) \qquad \forall x \in \overline{\tau},$$

co wynika wprost z definicji bazy nodalnej V^h i operatora interpolacji nodalnej π_h (por. definicje 13.4 i 13.5). Zauważmy, że twierdzenia Sobolewa o włożeniu (ang. *Sobolev embedding theorem*), zob. twierdzenie 16.3, wynika, że dla $d \leqslant 3$ zachodzi $H^2(\Omega) \subset C(\Omega)$. Z twierdzenia 16.5 (biorąc $m = 0, 1$ i $l + 1 = 2$) otrzymujemy:

$$|u - \pi_h u|^2_{H^m(\Omega)} = \sum_{\tau \in T_h} |u - \pi_h u|^2_{H^m(\tau)} \leqslant C \sum_{\tau \in T_h} h^{4-2m} |u|^2_{H^2(\tau)} = C h^{4-2m} |u|^2_{H^2(\Omega)}, \quad m = 0, 1,$$

co kończy dowód. □

13.7. Zadania dyskretne i zbieżność

Dla rodziny triangulacji i przestrzeni ciągłych funkcji elementu skończonego $\{V^h\}_h$, zawierających funkcje zerujące się na brzegu, możemy wprowadzić zadanie dyskretne, a dokładniej rodzinę zadań dyskretnych (13.3), które mają jednoznaczne rozwiązania i są stabilne, tzn. wspólnie ograniczone (por. (13.4)).

Teraz możemy wykorzystać teorię ciągłego elementu skończonego, aby otrzymać zbieżność i oszacowanie błędu dla elementu liniowego, por. rozdział 12.1.2, ale również elementów wyższego rzędu:

Wniosek 13.3. *Załóżmy, że spełnione są założenia twierdzenia 13.4 o rodzinie przestrzeni elementu skończonego $\{V^h\}$ zawartych w $H^1_0(\Omega)$. Rozpatrzmy $u^* \in H^1_0(\Omega)$ rozwiązanie (13.5) i u^*_h rozwiązanie zadania dyskretnego (13.3) z formą dwuliniową z (13.5) i przestrzenią dyskretną $V_n = V^h$. Wtedy*

$$|u^* - u^*_h|_{H^1(\Omega)} \to 0 \qquad h \to 0,$$

a jeśli dodatkowo $u^* \in H^1_0(\Omega) \cap H^2(\Omega)$, to

$$|u^* - u^*_h|_{H^1(\Omega)} \leqslant C\, h\, |u^*|_{H^2(\Omega)}. \tag{13.8}$$

Dowód. Dla $u^* \in H^1_0(\Omega) \cap H^2(\Omega)$ oszacowanie błędu (13.8) wynika z lematu Céa (twierdzenie 13.2) i z twierdzenia 13.4.

Zauważmy, że z definicji przestrzeni $H^1_0(\Omega)$ wynika, że jeśli $u^* \in H^1_0(\Omega)$ to dla dowolnego $\hat{\epsilon} > 0$ istnieje $u_\epsilon \in C^\infty_0(\Omega)$ takie, że

$$|u^* - u_\epsilon|_{H^1(\Omega)} \leqslant \hat{\epsilon}.$$

Następnie z lematu Céa (twierdzenie 13.2), nierówności trójkąta i z oszacowania z twierdzenia 13.4 otrzymujemy:

$$\begin{aligned} |u^* - u^*_h|_{H^1(\Omega)} &\leqslant C|u^* - \pi_h u_\epsilon|_{H^1(\Omega)} \leqslant C\left[|u^* - u_\epsilon|_{H^1(\Omega)} + |\pi_h u_\epsilon - u_\epsilon|_{H^1(\Omega)}\right] \\ &\leqslant C\hat{\epsilon} + C\, C_1\, h * |u_\epsilon|_{H^2(\Omega)}, \end{aligned}$$

dla C stałej z lematu Céa i C_1 stałej z twierdzenia 13.4. Stąd wynika zbieżność u^*_h do u^* w $H^1_0(\Omega)$ dla $h \to 0$. □

Dla dowolnego obszaru wielokątnego (wielościennego) niech T_h będzie rodziną triangulacji trójkątnych, jak w twierdzeniu 13.4, i niech dla $p = 1, 2, 3, \ldots$

$$V^h_p = \{u \in C(\overline{\Omega}) : u_{|\tau} \in P_p(\tau),\ \forall \tau \in T_h; \quad u = 0 \quad \text{na} \quad \partial\Omega\}.$$

Przestrzeń V^h_p nazywamy ciągłą przestrzenią elementu liniowego dla $p = 1$, kwadratowego dla $p = 2$ i kubicznego dla $p = 3$. Wtedy:

Wniosek 13.4. *$V_p^h \subset H_0^1(\Omega)$ i jeśli u^* jest rozwiązaniem (13.5), a $u_{h,p}^*$ jest rozwiązaniem zadania dyskretnego (13.3) z przestrzenią $V^n = V_p^h$ $p = 1, 2, 3, \ldots$ dla formy dwuliniowej z (13.5), to*

$$|u^* - u_{h,p}^*|_{H^1(\Omega)} \leqslant C\, h\, |u^*|_{H^2(\Omega)}$$

o ile $u^* \in H^2(\Omega)$.

13.8. Zadania

Ćwiczenie 13.1. Udowodnij (13.6).

Ćwiczenie 13.2. Niech $f, f_k \in L^2(\Omega)$ dla $\Omega \subset \mathbb{R}^d$. Pokaż, że $\Psi(u) = \int_\Omega fu\, dx + \sum_{k=1}^d f_k \frac{\partial u}{\partial x_k}\, dx$ zdefiniowane dla $u \in H_0^1(\Omega)$ jest ograniczonym funkcjonałem liniowym na $H_0^1(\Omega)$.

Ćwiczenie 13.3. Niech $\Psi \in (H_0^1(\Omega))^*$ będzie funkcjonałem liniowy na $H_0^1(\Omega)$. Tu $\Omega \subset \mathbb{R}^d$. Pokaż, że istnieją $f, f_1, f_2 \in L^2(\Omega)$ dla $\Omega \subset \mathbb{R}^d$, takie, że $\Psi(u) = \int_\Omega fu\, dx + \sum_{k=1}^d f_k \frac{\partial u}{\partial x_k}\, dx$ dla dowolnego $u \in H_0^1(\Omega)$.

Ćwiczenie 13.4 (trick Nitsche'go)**.** Rozpatrzmy zadanie dualne do (13.5): znaleźć $\psi \in H_0^1(\Omega)$

$$a(v, \psi) = \int_\Omega fv\, dx \qquad \forall v \in H_0^1(\Omega).$$

Pokaż, że ma ono jednoznaczne rozwiązanie takie, że $|\psi|_{H^1(\Omega)} \leqslant C\|f\|_{L^2(\Omega)}$.

Dodatkowo zakładamy regularność rozwiązania dualnego (13.5): tzn., że dla dowolnego $f \in L^2(\Omega)$ zachodzi: $\psi \in H^2(\Omega)$ z $|\psi|_{H^2(\Omega)} \leqslant C\|f\|_{L^2(\Omega)}$ oraz, że $V_h \subset H_0^1(\Omega)$ taki, że jeśli $\psi \in H^2(\Omega)$ to $\inf_{v \in V^h} \|\psi - v\|_{H^1(\Omega)} \leqslant C_1 h |\psi\|_{H^1(\Omega)}$.

Pokaż, że biorąc $f = u^* - u_h^*$ dla u_h^* rozwiązania (13.3) otrzymamy:

$$\|u^* - u_h^*\|_{L^2(\Omega)} \leqslant Ch^2 |u^*|_{H^2(\Omega)}.$$

Ćwiczenie 13.5. Dla liniowej przestrzeni elementu skończonego V^h na kwadracie jednostkowym z rozdziału 12.1.2 pokaż, że

$$|u^* - u_h^*|_{H^1(\Omega)} \leqslant Ch |u^*|_{H^2(\Omega)},$$

gdzie u^* rozwiązanie (12.1).

Wskazówka. Wystarczy sprawdzić założenia wniosku 13.4.

Ćwiczenie 13.6. Uogólnij twierdzenie 13.4 tzn. pokaż, zakładając, że $P_p \subset \hat{P}$, a $u \in H^{p+1}(\Omega) \cap H_0^1(\Omega)$, że otrzymujemy

$$|u - \pi_h u|_{H^s(\Omega)} \leqslant Ch^{p+1-s} |u|_{H^{p+1}(\Omega)} \qquad s = 0, 1, \ldots, p.$$

Ćwiczenie 13.7. Uogólnij wniosek 13.4, tzn. pokaż, że przy założeniach wniosku, jeśli dodatkowo $u^* \in H^{p+1}(\Omega)$, to

$$|u^* - u_{h,p}^*|_{H^1(\Omega)} \leqslant C\, h^p\, |u^*|_{H^{p+1}(\Omega)}.$$

Wskazówka. Wykorzystaj wynik poprzedniego zadania.

14. Metody numeryczne rozwiązywania równań parabolicznych drugiego rzędu

W tym rozdziale zajmiemy się metodami rozwiązywania równań ewolucyjnych drugiego rzędu, czyli równaniami parabolicznymi (2.9).

Takie równania możemy przedstawić abstrakcyjnie jako zadanie znalezienia funkcji $u : [0,T] \to V$ spełniającej:

$$\begin{aligned} \frac{du}{dt} - Lu(t) &= f(t) \qquad t \in (0,T] && (14.1)\\ u(0) &= u_0 \in V, && (14.2) \end{aligned}$$

dla V przestrzeni Hilberta, czy - ogólniej - Banacha, L operatora liniowego określonego dla elementów z V i danej funkcji f określonej na $(0,T]$ o wartościach w V^* z przestrzeni sprzężonej. Dane równanie możemy zdyskretyzować po przestrzeni wprowadzając przestrzeń dyskretną skończenie wymiarową V^h aproksymującą V, operator L_h określony na V^h aproksymujący L, funkcję $u_{0,h}$ aproksymująca u_0, oraz $f_h : (0,T) \to V_h$ przybliżenie f. V^h może być zbudowane metodą różnic skończonych albo elementu skończonego, lub jeszcze inną metodą dyskretyzacji np. metodą spektralną nie omawianą w tym skrypcie, por. np. [26].

Za aproksymację problemu wyjściowego możemy przyjąć $u_h : [0,T] \to V^h$ rozwiązanie następującego układu równań zwyczajnych pierwszego rzędu z warunkami początkowymi:

$$\begin{aligned} \frac{du_h}{dt} - L_h u_h(t) &= f_h(t) \qquad t \in (0,T] && (14.3)\\ u_h(0) &= u_{0,h} \in V^h. \end{aligned}$$

Następnie ten układ możemy rozwiązać przy pomocy jednego ze schematów dla równań zwyczajnych opisanych w pierwszych rozdziałach niniejszego skryptu (por. rozdziały 3-6).

Jeśli chodzi o analizę takich schematów, to stosuje się dwa podejścia: pierwszym jest szacowanie w odpowiedniej normie $u(t) - u_h(t)$ (jeśli $V_h \not\subset V$ to z odpowiednim przedłużeniem $r_h u_h \in V$), a następnie skorzystanie z ogólnej teorii zbieżności dla schematów dla równań zwyczajnych zastosowanych do rozwiązania (14.3). Drugim podejściem jest konstrukcja schematu całkowicie dyskretnego. Np. dyskretyzujemy (14.3) po czasie jakimś schematem dla zadań początkowych dla równań zwyczajnych, np. którymś ze schematów Eulera, czy trapezów lub jakimś schematem wyższego rzędu, a następnie przeprowadzamy analizę tak powstałego schematu dyskretnego.

Jeśli dyskretyzujemy równanie po zmiennej przestrzennej metodą różnic skończonych, a następnie po czasie - za pomocą jakiegoś schematu ze stałym krokiem całkowania, to tak otrzymany schemat możemy analizować korzystając z ogólnej teorii schematów różnicowych Laxa (por. rozdział 8.1).

Jeśli dyskretyzujemy wyjściowe zadanie paraboliczne przy pomocy metody elementu skończonego, to częściej - choć nie zawsze - do analizy stosuje się podejście pierwsze; tzn.: najpierw badamy błąd w odpowiedniej normie przestrzeni Sobolewa pomiędzy u a u_h rozwiązaniem (14.3), a następnie układ (14.3) przepisujemy jako układ równań zwyczajnych na współczynniki rozwiązania w ustalonej bazie V^h.

14.1. Schematy różnicowe dla modelowych równań parabolicznych

W tym rozdziale przedstawimy kilka możliwych schematów dla modelowych równań parabolicznych w jednym i dwóch wymiarach.

14.1.1. Przypadek jednowymiarowy

Rozpatrzmy następujące równanie paraboliczne z jednorodnymi warunkami brzegowymi: należy znaleźć funkcję u określoną na $[0,T]$ taką, że

$$\begin{aligned} u_t - u_{xx} + cu &= f(t,x) \qquad t \in (0,T] \quad x \in \Omega = (0,l) \\ u(t,0) = u(t,l) &= 0 \qquad t \in [0,T] \\ u(0,x) &= u_0(x) \qquad x \in (0,l) \end{aligned} \tag{14.4}$$

dla f danej funkcji ciągłej określonej na $[0,T]\times(0,l)$ i ciągłej funkcji u_0 określonej na $(0,l)$ i c stałej nieujemnej.

Wprowadzając siatkę jednorodną w obszarze Ω: $\overline{\Omega}_h = \{x_k\}_{k=0}^{M}$ z $x_k = k*h$ dla $h = l/M$ ($\Omega_h = \{x_k\}_{k=1}^{M-1}$) i zastępując operator

$-\frac{\partial^2}{\partial x^2} + c$ przez operator siatkowy dyskretny $-\overline{\partial}\partial_h + c$, por. (7.3), dobrze określony dla funkcji dyskretnych na siatce, otrzymujemy układ równań zwyczajnych, którego rozwiązanie powinno aproksymować (14.4):

$$\begin{aligned} \frac{du_h}{dt}(t,x) - \overline{\partial}\partial_h u_h(t,x) + cu_h(t,x) &= f(t,x) \qquad t \in (0,T], \quad x \in \Omega_h, \\ u_h(t,x_0) = u_h(t,x_M) &= 0 \qquad t \in [0,T], \\ u_h(0,x) &= u_0(x) \qquad x \in \Omega_h. \end{aligned} \tag{14.5}$$

Jeśli wprowadzimy dyskretne kroki czasowe na odcinku $[0,T]$: $t_n = n*\tau$ dla $n = 0,\ldots,N$ i $\tau = T/N$, to układ równań zwyczajnych (14.5) możemy zdyskredytować po czasie używając któregoś ze schematów ze stałym krokiem dla równań zwyczajnych. Czyli np.: otwarty schemat Eulera daje nam schemat różnicowy (zwany otwartym schematem Eulera dla równania parabolicznego) polegający na tym, że należy znaleźć $\{u_k^n\}_{k=1,\ldots,M;n=0,\ldots,N}$ takie, że

$$\begin{aligned} \frac{u_k^n - u_k^{n-1}}{\tau} + \frac{-u_{k-1}^{n-1} + (2 + c*h^2)*u_k^{n-1} - u_{k-1}^{n-1}}{h^2} &= f_k^{n-1} \qquad 0 < k < M,\ n = 1,\ldots,N \\ u_0^n = u_M^n &= 0 \qquad n = 0,\ldots,N \\ u_k^0 &= u_0(x_k) \qquad k = 1,\ldots,M-1 \end{aligned} \tag{14.6}$$

dla $f_k^n = f(t_n,x_k)$. Tutaj przyjęliśmy oznaczenie, że szukane przybliżenie $u_h(t_n,x_k)$ oznaczamy przez u_k^n.

Powyższy schemat możemy potraktować jako schemat różnicowy na siatce dyskretnej $\Omega_{\tau,h} = \{(t_n,x_k)\}_{n,k}$ z parametrem siatki $\max(\tau,h)$ dla obszaru wyjściowego $\Omega_T = (0,T]\times\Omega$ i operatorem różnicowym (siatkowym) $L_{\tau,h} = \partial_\tau - \overline{\partial}\partial_h + cI$, por. (7.2), przybliżającym na $\Omega_{\tau,h}$ operator paraboliczny $L = \frac{\partial}{\partial t} - \frac{\partial^2}{\partial x^2} + cI$. Następnie można pokazać, że jeśli u rozwiązanie wyjściowego problemu jest dostatecznie gładkie, to otrzymujemy: $|L_{\tau,h}u(t_n,x_k) - Lu(t_n,x_k)| = O(\tau + h^2) = O(\max(\tau,h^2))$, czyli rząd aproksymacji schematu wynosi jeden (por. definicję 8.4). A bardziej szczegółowo - rząd aproksymacji schematu wynosi jeden względem τ, a względem h wynosi dwa.

Można pokazać stabilność operatora różnicowego $L_{\tau,h}$ w odpowiednich normach dyskretnych (por. definicję 8.5), tzn. odpowiednia norma dyskretna u_h jest nie większa niż stała niezależna od h,τ pomnożona przez odpowiednie normy dyskretne $\{f_k^n\}_{n,k}$ i $\{u_k^0\}_k$.

Przypomnijmy, że jeśli schemat różnicowy jest stabilny i posiada odpowiedni rząd aproksymacji schematu, to jest zbieżny z odpowiednim rzędem, por. rozdział 8.1.

Niestety - stabilność jest tylko warunkowa, tzn. tylko dla h i τ spełniających odpowiedni warunek. Mówimy wtedy, że schemat jest stabilny warunkowo. Z praktycznego punktu widzenia lepiej byłoby gdyby schemat był stabilny absolutnie, tzn. dla dowolnej pary $\tau > 0, h > 0$.

Analogicznie możemy wprowadzić zamknięty schemat Eulera dla modelowego zadania parabolicznego stosując zamknięty schemat Eulera dla równań zwyczajnych do dyskretyzacji po czasie (14.5):

$$\begin{aligned} \frac{u_k^n - u_k^{n-1}}{\tau} + \frac{-u_{k-1}^n + (2 + c * h^2) * u_k^n - u_{k-1}^n}{h^2} &= f_k^n \qquad 0 < k < M, \quad n = 1, 2, \ldots, N \\ u_0^n = u_M^n &= 0 \qquad n = 0, \ldots, N \\ u_k^0 &= u_0(x_k) \qquad k = 1, \ldots, M-1 \end{aligned} \tag{14.7}$$

Rząd lokalnego błędu aproksymacji zamkniętego schematu Eulera jest taki sam jak otwartego schematu Eulera, ale dla $c > 0$ schemat ten jest absolutnie stabilny w dyskretnej normie maksimum. Kolejny schemat Cranka-Nicholson otrzymany po zastosowaniu schematu trapezów, por. (4.7), do (14.5):

$$\begin{aligned} &\frac{u_k^n - u_k^{n-1}}{\tau} + 0.5 * \left(\frac{-u_{k-1}^{n-1} + (2 + c\,h^2) * u_k^{n-1} - u_{k-1}^{n-1}}{h^2} + \frac{-u_{k-1}^n + (2 + c\,h^2) * u_k^n - u_{k-1}^n}{h^2} \right) \\ &= 0.5 * (f_k^{n-1} + f_k^n) \qquad 0 < k < M, \quad n = 1, 2, \ldots, N \\ &u_0^n = u_M^n = 0 \qquad n = 0, \ldots, N \\ &u_k^0 = u_0(x_k) \qquad k = 1, \ldots, M-1 \end{aligned}$$

Można pokazać, że lokalny błąd aproksymacji tego schematu jest jak $O(\tau^2 + h^2)$.

W przypadku zamkniętego schematu Eulera i schematu Cranka-Nicholson można pokazać ich bezwarunkową stabilność dla $c \geqslant 0$ w specjalnie dobranych normach dyskretnych.

14.1.2. Przypadek dwuwymiarowy na kwadracie

W tym rozdziale zajmiemy się ze względu na prostotę prezentacji modelowym równaniem parabolicznym z jednorodnymi warunkami brzegowymi na kwadracie $\Omega = (0,1)^2$. Chcemy znaleźć funkcję u określoną na $[0,T]$ taką, że

$$\begin{aligned} u_t - \triangle u + cu &= f(t,x) \qquad t \in (0,T] \quad x \in \Omega = (0,1)^2 \\ u(t,s) &= 0 \qquad t \in [0,T] \quad s \in \partial\Omega \\ u(0,x) &= u_0(x) \qquad x \in \Omega \end{aligned} \tag{14.8}$$

gdzie f - to dana funkcja ciągła określona na $(0,T] \times (0,1)^2$, u_0 - to funkcja ciągła określona na $[0,1]^2$, a c - to stała nieujemna.

Wprowadzając siatkę jednorodną w obszarze Ω jak w rozdziale 7.2: $\Omega_h, \overline{\Omega}_h, \partial\Omega_h$ dla $h = 1/M$, i zastępując operator $-\triangle + c$ przez operator siatkowy dyskretny $-\sum_{s=1}^{2} \overline{\partial}\partial_{s,h} + c$ por. (7.9) dobrze określony dla funkcji dyskretnych na jednorodnej siatce, otrzymujemy układ równań zwyczajnych, którego rozwiązanie powinno aproksymować (14.8):

$$\begin{aligned} \frac{du_h}{dt}(t,x) + (-\sum_{s=1}^{2} \overline{\partial}\partial_{s,h} + c)u_h(t,x) &= f(t,x) \qquad x \in \Omega_h, \quad t \in (0,T] \\ u_h(t,x_0) &= 0 \qquad t \in [0,T] \quad s \in \partial\Omega_h \\ u_h(0,x) &= u_0(x) \qquad x \in \Omega_h \end{aligned} \tag{14.9}$$

Tak jak w przypadku jednowymiarowym (por. rozdział 14.1.1), wprowadzamy dyskretną siatkę po zmiennej czasowej z krokiem τ na odcinku $[0,T]$: $t_n = n * \tau$ dla $n = 0,\ldots,N$ i $\tau = T/N$ i otrzymujemy dyskretyzację układu równań zwyczajnych (14.9) używając któregoś ze schematów ze stałym krokiem dla równań zwyczajnych.

Otwarty schemat Eulera daje nam następujący schemat polegający na znalezieniu $\{u^n_{k,l}\}$ takiego, że:

$$\begin{aligned} \frac{u^n_{k,l} - u^{n-1}_{k,l}}{\tau} + (-\sum_{s=1}^{2} \overline{\partial}\partial_{s,h} + c)u^{n-1}_{k,l} &= f^{n-1}_{k,l} \qquad 0 < k,l < M,\ n = 1,\ldots,N \\ u^n_{k,l} &= 0 \qquad n = 0,\ldots,N \quad k,l = 0, M \\ u^0_{k,l} &= u_0(k*h, l*h) \qquad k,l = 1,\ldots,M-1 \end{aligned} \tag{14.10}$$

Przybliżenie $u_h(t_n,(k*h,l*h))$ oznaczamy przez $u^n_{k,l}$.

W szczególności otrzymujemy

$$(-\sum_{s=1}^{2} \overline{\partial}\partial_{s,h} + c)u^n_{k,l} = \frac{1}{h^2}(-u^n_{k,l-1} - u^n_{k-1,l} + 4u^n_{k,l} - u^n_{k+1,l} - u^n_{k,l+1}) + cu^n_{k,l}$$

Analogicznie możemy zdefiniować schemat zamknięty Eulera lub schemat Cranka-Nicholson, czyli schemat trapezów zastosowany do (14.9).

14.2. Metoda elementu skończonego dla modelowych zadań

14.2.1. Przypadek jednowymiarowy

Rozpatrzmy ponownie jednowymiarowe modelowe zadanie (14.4). Jego słabe sformułowanie wprowadzamy analogicznie jak w rozdziale 11. Mnożąc równanie paraboliczne (14.4) przez funkcję testową z $C_0^\infty(0,l)$, całkując po $(0,l)$ i stosując wzór na całkowanie przez części otrzymujemy równanie:

$$(u_t,\phi)_{L^2(0,l)} + \left(\frac{du}{dx},\frac{d\phi}{dx}\right)_{L^2(0,l)} + c(u,\phi)_{L^2(0,l)} = (f(t,\cdot),\phi)_{L^2(0,l)} \qquad \forall \phi \in C_0^\infty(0,l)$$

z warunkiem początkowym

$$(u(0),\phi)_{L^2(0,l)} = (u_0,\phi)_{L^2(0,l)} \qquad \forall \phi \in C_0^\infty(0,l).$$

Korzystając z tego, że $H_0^1(0,l)$ jest domknięciem $C_0^\infty(0,l)$ w normie H^1 można pokazać, że powyższe równanie jest równoważne znalezieniu funkcji $u : [0,T] \to H_0^1(\Omega)$ takiej, że

$$\begin{aligned} \frac{d}{dt}(u,v)_{L^2(0,l)} + \left(\frac{du}{dx},\frac{dv}{dx}\right)_{L^2(0,l)} + c(u,v)_{L^2(0,l)} &= (f(t,\cdot),v)_{L^2(0,l)} \quad \forall v \in H_0^1(0,l), \\ (u(0),v)_{L^2(0,l)} &= (u_0,v)_{L^2(0,l)} \qquad \forall v \in H_0^1(0,l) \end{aligned} \tag{14.11}$$

dla $0 < t \leqslant T$, co jest słabym (wariacyjnym) sformułowaniem (14.4), które stanowi wyjście do konstrukcji dyskretyzacji równania parabolicznego za pomocą metody elementu skończonego. Niech $T_h([0,l]) = \{[x_k, x_{k+1}\}$ będzie triangulacją równomierną $[0,l]$ zdefiniowaną jak w rozdziale 11.1.2, tzn. $x_k = k*h$ dla $h = l/N$ i niech V^h będzie przestrzenią funkcji ciągłych kawałkami liniowych (tzn. liniowych na elementach $[x_k, x_{k+1}]$) zerujących się w końcach odcinka $[0,l]$. Oczywiście zachodzi $V^h \subset H_0^1(0,l)$.

Wtedy możemy zdefiniować dyskretyzację po przestrzeni zadania (14.11).

Znajdź funkcję $u_h : [0,T] \to V^h$ taką, że dla $0 < t \leqslant T$ i dowolnego $v_h \in V^h$ zachodzi:

$$\begin{aligned} \frac{d}{dt}(u_h, v_h)_{L^2(0,l)} + \left(\frac{du_h}{dx}, \frac{dv_h}{dx}\right)_{L^2(0,l)} + c(u_h, v_h)_{L^2(0,l)} &= (f(t,\cdot), v_h)_{L^2(0,l)} \\ (u_h(0), v_h)_{L^2(0,l)} &= (u_0, v_h)_{L^2(0,l)} \end{aligned} \tag{14.12}$$

Biorąc bazę nodalną tej przestrzeni $(\phi_k)_{k=1}^{N-1}$ (por. (11.3)) i rysunek 11.1 na str. 98, otrzymujemy $u_h = \sum_{k=1}^{N-1} u_k \phi_k$ i

$$\begin{aligned} M_h \frac{d}{dt}\vec{u} + (A_h + cM_h)\vec{u} &= \vec{f}(t) \\ M_h \vec{u}(0) &= \vec{u}_{0,h} \end{aligned}$$

dla $\vec{u} = (u_k)_k$, $M_h = ((\phi_k, \phi_l)_{L^2(0,l)})_{k,l}$, $A_h = ((\frac{d\phi_k}{dx}, \frac{d\phi_l}{dx})_{L^2(0,l)})_{k,l}$, $\vec{u}_{0,h} = ((u_0, \phi_k)_{L^2(0,l)})_k$ i wektora prawej strony $\vec{f} = (f_1, \ldots, f_{N-1})^T$ dla $f_k(t) = (f(t), \phi_k)_{L^2(0,l)}$. Z tego otrzymujemy

$$\begin{aligned} \frac{d}{dt}\vec{u} + (M_h^{-1}A_h + cI)\vec{u} &= M_h^{-1}\vec{f}(t) =: \vec{g}(t) \\ \vec{u}(0) &= M_h^{-1}\vec{u}_{0,h} := \vec{u}_0. \end{aligned}$$

Proszę zauważyć, że jest to układ równań zwyczajnych liniowych z warunkiem początkowym, więc ma jednoznaczne rozwiązanie na $[0,T]$, co wynika z ogólnej teorii równań różniczkowych zwyczajnych, por. np. rozdział 3.2 lub [23]. Do powyższego układu równań możemy zastosować dowolny schemat rozwiązywania równań zadania początkowego dla równań zwyczajnych.

Macierze M_h i A_h są symetryczne i dodatnio określone.

Można pokazać, że macierz $M_h^{-1}A_h$ ma wartości własne ujemne o module od jeden do rzędu h^{-2}, czyli bardzo dużym module dla małych h. Zatem dla $c = 0$ układ równań zwyczajnych jest sztywny zgodnie z definicją z rozdziału 6. Należy tu stosować schematy całkowania równań zwyczajnych stosowne do zadań sztywnych.

14.2.2. Przypadek dwuwymiarowy

Rozpatrzmy dwuwymiarowe modelowe zadanie na dowolnym obszarze wielokątnym na płaszczyźnie Ω, czyli zastępując kwadrat przez Ω w (14.8). Jego słabe sformułowanie otrzymujemy analogicznie jak w rozdziale 11, lub w przypadku jednowymiarowym (por. rozdział 14.2.1). Mnożąc równanie paraboliczne z (14.8) przez funkcję testową z $C_0^\infty(\Omega)$, całkując po Ω i stosując wzory Greene'a otrzymujemy:

$$(u_t, \phi)_{L^2(\Omega)} + a(u,v) = (f(t,\cdot), \phi)_{L^2(\Omega)} \qquad \forall \phi \in C_0^\infty(\Omega)$$

dla $0 < t \leqslant T$, $a(u,v) = (\nabla u, \nabla \phi)_{L^2(\Omega)} + c(u,\phi)_{L^2(\Omega)}$ oraz u spełnia warunek początkowy $(u(0), \phi)_{L^2(\Omega)} = (u_0, \phi)_{L^2(\Omega)}$.

Jak w rozdziale 14.2.1 otrzymujemy, że powyższe równanie jest równoważne znalezieniu funkcji $u : [0,T] \to H_0^1(\Omega)$ takiej, że

$$\begin{aligned} \frac{d}{dt}(u,v)_{L^2(\Omega)} + a(u,v) &= (f(t,\cdot), v)_{L^2(\Omega)} \quad \forall v \in H_0^1(\Omega), \\ (u(0), v)_{L^2(\Omega)} &= (u_0, v)_{L^2(\Omega)} \quad \forall v \in H_0^1(\Omega) \end{aligned} \tag{14.13}$$

dla $0 < t \leqslant T$, co jest słabym, wariacyjnym sformułowaniem (14.8), które stanowi wyjście do konstrukcji dyskretyzacji równania parabolicznego za pomocą metody elementu skończonego.

Rozpatrzmy $T_h(\Omega)$ triangulacje równomierną Ω, złożoną z przystających trójkątów, zdefiniowaną jak w rozdziale 12 i V^h - przestrzeń funkcji ciągłych kawałkami liniowych na tej triangulacji, zerujących się na brzegu, czyli przestrzenią liniowego elementu skończonego (por. rozdział 12).

Dyskretyzację po przestrzeni zadania (14.13) definiujemy: znajdź funkcję $u_h : [0,T] \to V^h$ taką, że dla $0 < t \leqslant T$ i dowolnego $v_h \in V^h$:

$$\begin{aligned} \frac{d}{dt}(u_h, v_h)_{L^2(\Omega)} + a(u_h, v_h) &= (f(t,\cdot), v_h)_{L^2(\Omega)} \\ (u_h(0), v_h)_{L^2(\Omega)} &= (u_0, v_h)_{L^2(\Omega)}. \end{aligned} \tag{14.14}$$

Otrzymaliśmy zatem ponownie układ równań zwyczajnych liniowych z warunkiem początkowym, który po wprowadzeniu standardowej bazy daszkowej $\{\phi_{kl}\}$ dla V^h, por. (12.2), możemy przepisać jako zadanie początkowe na funkcje-współczynniki $\alpha_{k,l}(t)$ takie, że $u_h(t) = \sum_{k,l} \alpha_{k,l}(t)\phi_{kl}$. Następnie to zadanie początkowe możemy rozwiązać za pomocą jakiegoś schematu, np. otwartego lub zamkniętego schematu Eulera, lub schematu trapezów. Okazuje się, że - tak samo jak w przypadku jednowymiarowym - dla $c = 0$ powstające układy równań zwyczajnych są sztywne. Dlatego w praktyce stosuje się odpowiednie schematy dla zadań sztywnych.

14.3. Zadania

Ćwiczenie 14.1. Zbadaj rzędy błędów aproksymacji otwartego schematu Eulera (14.6) i zamkniętego schematu Eulera (14.7) dla dyskretyzacji modelowego problemu jednowymiarowego w dyskretnej normie maksimum przyjmując, że rozwiązanie jest dostatecznie gładkie. Ustal, jaka minimalna gładkość rozwiązania jest konieczna, tzn. znajdź najmniejsze r takie, że jeśli rozwiązanie $u \in C^r$, to rząd aproksymacji schematu jest możliwie duży.

Ćwiczenie 14.2. Zbadaj stabilność zamkniętego schematu Eulera (14.7) dla dyskretyzacji modelowego problemu jednowymiarowego w dyskretnej normie maksimum dla $c > 0$. Wywnioskuj zbieżność dyskretną schematu w tejże normie.

Wskazówka. Zastosuj twierdzenie 9.1, a do wykazania zbieżności zastosuj ogólną teorię zbieżności schematów różnicowych Laxa z rozdziału 8.

Ćwiczenie 14.3. Zbadaj rząd błędu aproksymacji schematu Cranka-Nicholson dla $c > 0$ dla dyskretyzacji modelowego problemu jednowymiarowego w dyskretnej normie maksimum przyjmując, że rozwiązanie jest dostatecznie gładkie. Ustal, jaka minimalna gładkość rozwiązania jest konieczna, aby schemat miał ten rząd. Zbadaj stabilność tego schematu w dyskretnej normie maksimum: czy jest warunkowa, czy bezwarunkowa? Zbadaj zbieżność w dyskretnej normie maksimum.

Ćwiczenie 14.4. Rozpatrzmy równanie paraboliczne dla $\Omega = (0,1)^2$.

— Zapisz (14.14) w postaci liniowego zadania początkowego:

$$M_h \frac{d\vec{u}}{dt}(t) + A_h \vec{u}(t) = \vec{f}(t), \qquad \vec{u}(0) = \vec{u}_0$$

dla $\vec{u}(t) = \{c_{k,l}(t)\}_{k,l}$ współczynników $u_h(t)$ w bazie daszkowej V^h (por. (12.2)) i M_h, A_h macierzy stałych.

— Wypisz wzory na otwarty i zamknięty schemat Eulera zastosowany do tego zadania początkowego.

15. Metody numeryczne rozwiązywania równań hiperbolicznych pierwszego rzędu

W tym rozdziale zajmiemy się metodami rozwiązywania równań hiperbolicznych pierwszego rzędu (por. rozdział 2.2.2). Przedstawimy konstrukcję kilku otwartych schematów różnicowych oraz podamy ideę zbieżności schematów za [26].

Konstrukcję schematów różnicowych przedstawimy dla modelowych równań, tzn. równań liniowych skalarnych, czyli będziemy szukali przybliżeń funkcji $u = u(t, x)$ takiej, że

$$u_t + a(t, x)u_x = b(t, x), \tag{15.1}$$

Zazwyczaj rozwiązania będą spełniały też warunek początkowy $u(0, x) = u_0(x)$, przy czym najczęściej będziemy zakładać dla prostoty prezentacji, że a jest stałą, a $b = 0$.

Pokażemy jak stosować te schematy dla równań nieliniowych i układów równań liniowych postaci:

$$\vec{u}_t + A\vec{u}_x = 0, \tag{15.2}$$

gdzie A - to stała macierz $m \times m$ diagonalizowalna w jakiejś bazie (ponieważ jest to układ hiperboliczny).

15.1. Schematy różnicowe dla równania skalarnego

W tym rozdziale zajmiemy się schematami różnicowymi dla równania skalarnego (15.1). Zakładamy, że u spełnia dany warunek początkowy:

$$u(0, x) = u_0(x) \qquad x \in \mathbb{R}.$$

Rozpatrzmy siatkę równomierną na półpłaszczyźnie $[0, \infty) \times \mathbb{R}$ z krokiem przestrzennym $h > 0$ i czasowym $\tau > 0$:

$$(t_n, x_k) \qquad n \in \mathbb{N} \quad k \in \mathbb{Z}$$

dla

$$t_n = n * \tau, \qquad x_k = k * h.$$

Możemy wtedy zdefiniować najprostszy otwarty schemat w sposób następujący:

$$\tau^{-1}(u_k^{n+1} - u_k^n) + h^{-1}a(u_k^n - u_{k-1}^n) = 0,$$

lub

$$\tau^{-1}(u_k^{n+1} - u_k^n) + h^{-1}a(u_{k+1}^n - u_k^n) = 0.$$

W tym rozdziale zakładamy, że spełniony jest warunek początkowy $u_k^0 = u_0(x_k)$ dla $k \in \mathbb{Z}$.

Oba schematy są otwarte. Można postawić pytanie: który ma lepsze własności? Okazuje się, że stabilność tych schematów zależy od znaku parametru a.

Schemat upwind definiujemy jako

$$\text{(Upwind)} \qquad \tau^{-1}(u_k^{n+1} - u_n^k) = \begin{cases} -h^{-1}a(u_{k+1}^n - u_k^n) & a < 0 \\ -h^{-1}a(u_k^n - u_{k-1}^n) & a > 0 \end{cases} \tag{15.3}$$

lub równoważnie biorąc $\lambda = \frac{\tau}{h}$:

$$\text{(Upwind)} \qquad (u_k^{n+1} - u_n^k) + a\frac{\lambda}{2}(u_{k+1}^n - u_{k-1}^n) = 0.5\lambda|a|(u_{k+1}^n - 2u_k^n + u_{k-1}^n) \tag{15.4}$$

Jeśli wprost dyskretyzujemy pochodną po przestrzeni za pomocą różnicy centralnej, to otrzymujemy następujący schemat:

$$\tau^{-1}(u_k^{n+1} - u_n^k) + a\frac{u_{k+1}^n - u_{k+1}^n}{2h} = 0$$

czyli

$$u_k^{n+1} = u_k^n - a\frac{\lambda}{2}(u_{k+1}^n - u_{k+1}^n). \tag{15.5}$$

Schemat ten niestety okazuje się być **niestabilnym** (wg definicji, która pojawi się później). Różni się on od poprzedniego schematu upwind (15.3) brakiem dodatkowego członu

$$0.5\lambda|a|(u_{k+1}^n - 2u_k^n + u_{k-1}^n) = \tau * h * 0.5|a|\frac{u_{k+1}^n - 2u_k^n + u_{k-1}^n}{h^2},$$

który aproksymuje, por. (7.3) i (7.4):

$$\tau * h * 0.5 * |a|\frac{\partial^2 u}{\partial x^2},$$

Ten człon można traktować jako sztuczną numeryczną lepkość (ang. *numerical dissipation or artificial viscosity*) dodaną do niestabilnego schematu (15.5), dzięki której schemat upwind (15.3) jest stabilny.

Wyprowadzimy teraz kilka kolejnych otwartych schematów.

Rozważmy teraz schemat Laxa-Friedrichsa, w którym u_k^n w niestabilnym schemacie (15.5) zastępujemy średnią z u_{k+1}^n i u_{k-1}^n i otrzymujemy:

$$\text{(Lax} - \text{Friedrichs)} \qquad u_k^{n+1} = 0.5(u_{k+1}^n + u_{k-1}^n) - a\frac{\lambda}{2}(u_{k+1}^n - u_{k+1}^n). \tag{15.6}$$

Kolejny schemat Laxa-Wendroffa wyprowadza się z rozwinięcia rozwiązania w momencie $t = t_n$ w szereg Taylora:

$$u(t_n + \tau, x_k) = u(t_n, x_k) + \tau\frac{\partial u}{\partial t}(t_n, x_k) + 0.5\tau^2\frac{\partial^2 u}{\partial t^2}(t_n, x_k) + O(\tau^3).$$

Korzystamy następnie z równania:

$$\frac{\partial u}{\partial t} = -a\frac{\partial u}{\partial x}$$

i kolejnego równania otrzymanego z poprzedniego:

$$\frac{\partial^2 u}{\partial t^2} = a^2\frac{\partial^2 u}{\partial x^2}.$$

W tych równaniach zastępujemy pochodną po przestrzeni ilorazem różnicowym centralnym, por. (4.2):

$$\frac{\partial u}{\partial x}(t, x) \approx 0.5h^{-1}(u(t, x + h) - u(t, x - h)),$$

a drugą pochodną po przestrzeni jej przybliżeniem różnicowym na trzech punktach, por. (7.3) i (7.4):

$$\frac{\partial^2 u}{\partial x^2}(t, x) \approx 0.5h^{-2}(u(t, x - h) - 2 * u(t, x) + u(t, x + h))$$

i w końcu otrzymujemy schemat Laxa-Wendroffa:

$$(\text{Lax} - \text{Wendroff}) \qquad u_k^{n+1} = u_k^n - a\,0.5\lambda(u_{k+1}^n - u_{k-1}^n) + 0.5a^2\lambda^2(u_{k+1}^n - 2u_k^n + u_{k-1}^n). \quad (15.7)$$

Można też rozważać schematy wielopoziomowe ze względu na czas, np. schemat skoku żaby (leap-frog), w którym pochodną po czasie dyskretyzujemy przez pochodną centralną tak samo, jak pochodną po przestrzeni. Otrzymujemy wówczas schemat trzypoziomowy:

$$(\text{Leap} - \text{frog}) \qquad u_k^{n+1} = u_k^{n-1} - a\lambda(u_{k+1}^n - u_{k+1}^n). \quad (15.8)$$

15.2. Schematy dla równań nieliniowych lub układów równań

Zauważmy, że wszystkie dotąd rozważane dwupoziomowe schematy dla równań skalarnych, tzn. (15.7), (15.3), (15.5), można zapisać w zunifikowany sposób jako:

$$u_k^{n+1} = u_k^n - \lambda(H_{k+1/2}^n - H_{k-1/2}^n)$$

gdzie $H_{k+1/2}^n = H(u_k^n, u_{k+1}^n)$ jest określana jako numeryczny strumień (ang. *numerical flux*).

W ten sposób możemy schematy dla równania (15.1) łatwo przenieść na przypadek nieliniowych równań hiperbolicznych postaci:

$$u_t + \frac{\partial F(u)}{\partial x} = 0$$

gdzie F - to dana funkcja. $F(u)$ nazywamy strumieniem dla funkcji u. Wtedy każdy schemat można zapisać jako

$$u_k^{n+1} = u_k^n - \lambda(F_{k+1/2}^n - F_{k-1/2}^n)$$

przyjmując oznaczenie $F_{k+1/2}^n = H(F(u_k^n), F(u_k^{n+1}))$ z H numerycznym strumieniem wziętym z wyjściowego schematu. Zauważmy, że dla równania (15.1) zachodzi $F(u) = a * u$.

Również schematy te można zastosować do układu (15.2). Wtedy widzimy, że dla nieosobliwej macierzy C:

$$A = C\Lambda C^T,$$

gdzie $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_m)$ - to macierz diagonalna z wartościami własnymi A na diagonali (ponieważ jest to układ równań hiperbolicznych). Wtedy możemy zamienić zmienne: zamiast szukać wartości rozwiązania $U_k^n = \vec{u}(t_n, x_k)$ w punktach siatki, szukamy $W_k^n = CU_k^n$, czyli stosujemy schematy do równoważnego równania ($\vec{w} = C\vec{u}$):

$$\vec{w}_t + \Lambda\vec{w}_x = 0.$$

Proszę zauważyć, że to równanie jest układem m niezależnych równań hiperbolicznych (15.1), tzn.:

$$(w_j)_t + \lambda_j(w_j)_x = 0 \qquad j = 1, \ldots, m.$$

Zatem możemy zastosować np. schemat upwind lub inny - niezależnie do każdej składowej - i otrzymać przybliżone rozwiązanie $w_j(t_n, x_k)$.

Znając wartości W_k^n możemy wrócić do wyjściowych zmiennych i obliczyć U_k^n rozwiązując układ równań

$$CU_k^n = W_k^n,$$

czyli przybliżone rozwiązanie $\vec{u}(t_n, x_k)$.

W praktyce - w pierwszym kroku przeprowadzamy obliczenia wstępne rozwiązując numerycznie zadanie własne dla macierzy A, tzn. obliczając wartości i wektory własne A, czyli λ_j i kolumny C. Następnie stosujemy wybrany schemat obliczając wartości W_k^n dla punktów siatki (w praktyce musimy ograniczyć zakres k i n). Na koniec rozwiązujemy układ równań z macierzą C dla odpowiednich k i n otrzymując U_k^n.

15.3. Stabilność, zgodność i zbieżność schematów

Do schematów różnicowych dla równań hiperbolicznych stosuje się ogólną teorię zbieżności Laxa-Richmyera schematów różnicowych, analogiczną do teorii zbieżności z rozdziału 8.1, por. rozdział 14.2 w [26]. Aby uzyskać oszacowanie błędu w pewnej normie dyskretnej, należy wykazać odpowiedni rząd aproksymacji schematu i stabilność w tej normie.

Przy przyjętych powyżej oznaczeniach przyjmijmy, że $\mathbf{u}^n = \{u_k^n\}_{k\in\mathbb{Z}}$ i załóżmy, że dwupoziomowy (względem czasu) schemat różnicowy możemy opisać jako

$$\mathbf{u}^n = A_\tau(\mathbf{u}^{n-1}) \qquad n > 0, \tag{15.9}$$

lub w punkcie siatki x_k

$$u_k^n = A_\tau(u^{n-1}; k) \qquad n > 0.$$

ze znanym warunkiem początkowym $\mathbf{u}^0$, tzn. $u_k^0 = u_0(x_k)$. Stabilność w pewnej normie $\|\cdot\|_h$ na odcinku czasu $[0, T]$ oznacza, że istnieją stałe $\tau_0, C > 0$ takie, że dla czasów $t_n = n * \tau < T$ jeśli $0 < h, \tau < \tau_0$:

$$\|\mathbf{u}^n\|_h \leqslant C_T \|\mathbf{u}^0\|_h \qquad n > 0.$$

Często stosowaną normą jest norma będąca aproksymacją normy $L^1(\mathbb{R})$, czyli

$$\|\mathbf{u}^n\|_h = \sum_{k\in\mathbb{Z}} h|u_k^n|.$$

Jeśli istnieją stałe $\tau_0, \beta > 0$ takie, że dla czasów $t_n = n * \tau < T$ jeśli $0 < h, \tau < \tau_0$ zachodzi:

$$\|A_\tau \mathbf{u}\|_h \leqslant (1 + \beta\tau)\|\mathbf{u}\|_h \qquad \forall \mathbf{u},$$

to

$$\|\mathbf{u}^n\|_h \leqslant (1 + \beta\tau)^n \|\mathbf{u}^0\|_h \leqslant \exp(\beta T)\|\mathbf{u}^0\|_h \qquad n > 0$$

dla dowolnych n takich, że $t_n = n * \tau < T$, co oznacza stabilność schematu w normie $\|\cdot\|_h$.

Z kolei zgodność schematu (aproksymacja schematem wyjściowego zadania; ang. *consistency*) oznacza, że schemat dyskretny aproksymuje wyjściowe równanie, tzn.

$$(E_\tau(t))_k := \tau^{-1}|u(t + \tau, x_k) - A_\tau(\mathbf{u}(t); k)|$$

spełnia

$$\lim_{\tau\to 0} \|E_\tau(t)\|_h = 0$$

dla u rozwiązania wyjściowego równania, dowolnego $0 < h \leqslant \tau_0$ i $t > 0$. Tutaj $A_\tau(\mathbf{u}(t); k)$ jest zdefiniowane dla tego schematu jak $A_\tau(\mathbf{u}^n; k)$ zastępując u_k^n przez $u(t, x_k)$.

Jeśli dla pewnej stałej $C > 0$ i $0 < h, \tau \leqslant \tau_0$ zachodzi oszacowanie:

$$\|E_\tau(t)\|_h \leqslant C\,[\tau^{q_1} + h^{q_2}]$$

to powiemy, że rząd aproksymacji schematu wynosi q_1 po czasie i q_2 po przestrzeni. A jeśli zachodzi stała zależność τ od h np. liniowa $\tau = \kappa * h$ dla stałej κ, to mówimy, że rząd aproksymacji schematu wynosi $q = \min\{q_1, q_2\}$, co jest zgodne z definicją z rozdziału 8.1.

Jak już wiemy, teoria Laxa mówi, że stabilność i aproksymacja (zgodność) dają zbieżność. Tak jest też w tym przypadku, tzn. teoria Laxa-Richmyera (por. [27]) mówi, że schemat jest zbieżny:

$$\max_{0\leqslant n\leqslant T/\tau} \|u(t_n, \cdot) - \mathbf{u}^n\|_h \to 0$$

dla $h, \tau \to 0$ wtedy i tylko wtedy, jeśli jest stabilny i zgodny.

Przykład 15.1. Rozpatrzmy najprostszy schemat Laxa-Friedrichsa (15.6). Jeśli założymy, że

$$|a\lambda| \leqslant 1 \tag{15.10}$$

to otrzymujemy:

$$\begin{aligned} \|\mathbf{u}^{n+1}\|_h &= h\sum_k |u_j^{n+1}| \leqslant \frac{h}{2}\left[(1-\lambda a)\sum_j |u_{j+1}^n| + (1+\lambda a)\sum_j |u_{j-1}^n|\right] \\ &= \frac{1}{2}\left[(1-\lambda a)\|\mathbf{u}^n\|_h + (1+\lambda a)\|\mathbf{u}^n\|_h\right] = \|\mathbf{u}^n\|_h. \end{aligned}$$

czyli, że schemat jest stabilny warunkowo przy założeniu $|a\lambda| \leqslant 1$.

Analogicznie można pokazać, że przy założeniu, że spełniony jest warunek (15.10), schemat Laxa-Wendroffa (15.7) i schemat upwind (15.3) są stabilne. Widzimy, że schemat leap-frog (15.8) jest stabilny (w sensie analogicznej definicji odpowiedniej dla schematów trzypoziomowych) przy założeniu, że w warunku (15.10) zachodzi ostra nierówność.

Natomiast schemat (15.5) przy założeniach typu (15.10) nie jest stabilny.

Wykazanie, że wszystkie wymienione schematy są zgodne i zbadanie jakie mają rzędy pozostawiamy jako zadanie.

15.4. Metoda Fouriera badania stabilności

Metoda opisana w tym rozdziale służy badaniu stabilności schematów, choć - de facto - może tylko sprawdzić stabilność w sensie negatywnym. Tzn. metoda pozwala wykazać, że jakiś schemat nie jest stabilny.

Zakładamy, że rozpatrujemy schemat dwupoziomowy lub trzypoziomowy, np. dający się zapisać jako (15.9), i że szukamy jego rozwiązań w postaci

$$u_k^n = \gamma^n \exp(i\alpha k),$$

gdzie $\gamma \in \mathbb{C}$ i $\alpha \in \mathbb{R}$ są stałymi.

Metoda polega na wyznaczeniu warunków na te stałe w zależności od konkretnej postaci schematu. Jeśli się okaże, że istnieje rozwiązanie tej postaci z $|\gamma| > 1$, to oczywiście schemat stabilny być nie może, a jeśli wszystkie takie rozwiązania dla dowolnych α muszą spełniać $|\gamma| < 1$ - ewentualnie przy pewnych warunkach na τ i h - to schemat ma szanse być stabilnym, czy - inaczej: jest stabilnym w klasie rozwiązań tej postaci.

Podsumowując: wstawimy $u_k^n = \gamma^n e^{i\alpha k}$ do schematu i wyliczamy $\gamma(\alpha)$ jeśli $|\gamma(\alpha)| < 1$ dla dowolnego α, to schemat uważamy za stabilny w sensie opisanym powyżej.

Zbadanie stabilności schematów (15.7), (15.3), (15.8), (15.5) przy pomocy tej metody pozostawiamy jako zadanie.

15.5. Zadania

Ćwiczenie 15.1. Zbadaj przy pomocy metody Fouriera stabilność schematów:
- Laxa Wendroffa (15.7),
- schematu upwind (15.3),
- schematu *Leap-frog* (15.8),
- schematu opartego na różnicy centralnej (15.5)

Ćwiczenie 15.2 (Laboratoryjne)**.** Zaimplementuj w octave schemat Laxa-Wendroffa dla równania $au_x = u_t$ dla $a = 1, -1, 100, -100$, przyjmując, że znamy rozwiązanie początkowe na $[-1, 1]$ i warunki brzegowe $u(t, -1) = u(t, 1) = 0$. Zbadaj rząd metodą połowionych kroków, czyli dla h i $h/2$ policz błędy w ustalonym punkcie i ich stosunek.

16. Przestrzenie elementu skończonego, a aproksymacja w przestrzeniach Sobolewa

W tym rozdziale przedstawimy elementy teorii przestrzeni Sobolewa oraz kilka technicznych lematów potrzebnych do dowodów zbieżności metody elementu skończonego. Mimo, że przedstawimy tylko najmniej techniczne dowody odpowiednich lematów to, aby w pełni zrozumieć dowody, należałoby zapoznać się wcześniej z teorią przestrzeni Sobolewa, zob. np. [21].

Materiał w poniższym rozdziale wykracza poza materiał z wykładu.

16.1. Przestrzenie Sobolewa H^m

Poniżej podamy kilka faktów, dotyczących przestrzeni Sobolewa, potrzebnych do udowodnienia zbieżności metody elementu skończonego dla równania eliptycznego drugiego stopnia.

Najpierw zdefiniujmy przestrzenie Sobolewa $H^k(\Omega)$ dla $\Omega \subset \mathbb{R}^d$, por. [21].

Definicja 16.1. Rozpatrzmy $\Omega \subset \mathrm{R}^d$ obszar ograniczony, wtedy $H^m(\Omega)$ definiujemy jako przestrzeń funkcji z $L^2(\Omega)$, których słabe pochodne $\partial^\alpha u$ dla wszystkich $|\alpha| \leqslant m$ są w $L^2(\Omega)$. Iloczyn skalarny w H^m definiujemy jako

$$(u, v)_{H^m(\Omega)} = \sum_{|\alpha| \leqslant m} \partial^\alpha u \, \partial^\alpha v \, dx$$

z normą

$$\|u\|_{H^m(\Omega)} = \sqrt{\sum_{|\alpha| \leqslant m} |\partial^\alpha u|^2 \, dx}.$$

i półnormą

$$|u|_{H^m(\Omega)} = \sqrt{\sum_{|\alpha| = m} |\partial^\alpha u|^2 \, dx}.$$

Tutaj $\alpha = (\alpha_1, \ldots, \alpha_d)$ z $\alpha_j \in \mathbb{N}$ - to wielowskaźnik, $|\alpha| = \sum_{k=1}^d \alpha_k$ i

$$\partial^\alpha u = \frac{\partial^{|\alpha|} u}{\partial^{\alpha_1} \ldots \partial^{\alpha_d}}.$$

Można pokazać następujące twierdzenie:

Twierdzenie 16.1. *Rozpatrzmy $\Omega \subset \mathbb{R}^d$ otwarty obszar z kawałkami gładkim brzegiem i $m \geqslant 0$. Wtedy $C^\infty(\Omega) \cap H^m(\Omega)$ jest zbiorem gęstym w $H^m(\Omega)$.*

Proszę zauważyć, że to twierdzenie pozwala nam inaczej zdefiniować przestrzeń H^m jako domknięcie zbioru wszystkich funkcji gładkich, których norma $\|\cdot\|_{H^m(\Omega)}$ jest ograniczona.

Dodatkowo wprowadzamy:

Definicja 16.2. Niech $H_0^m(\Omega)$ będzie domknięciem w H^m przestrzeni C_0^∞, gdzie $C_0^\infty(\Omega)$ jest podprzestrzenią $C^\infty(\Omega)$ złożoną z funkcji o zwartym nośniku w Ω.

Zaznaczmy, że:

$$H_0^m(\Omega) \subset H^m(\Omega) \subset L^2(\Omega).$$

Zachodzą jeszcze następujące nierówności:

Stwierdzenie 16.1 (nierówność Friedrichsa)**.** *Jeśli Ω zawarty jest w jednostkowej kostce, to*

$$\|u\|_{L^2(\Omega)} \leqslant |u|_{H^1(\Omega)} \qquad \forall u \in H_0^1(\Omega).$$

Dowód w ogólności można znaleźć np. w [2], ale dla kostek w dwóch i trzech wymiarach dowód pozostawiamy jako zadanie.

Istnieje też następujące twierdzenie mówiące w jakim sensie możemy rozważać wartości funkcji z $H^1(\Omega)$ na brzegu tego obszaru.

Twierdzenie 16.2 (Twierdzenie o śladzie)**.** *Rozpatrzmy Ω ograniczony obszar o brzegu Lipschizowskim[a], wtedy istnieje ograniczony operator liniowy $\gamma : H^1(\Omega) \to L^2(\partial\Omega)$ i stała C:*

$$\|\gamma u\|_{L^2(\partial\Omega)} \leqslant C\|u\|_{H^1(\Omega)}. \qquad \forall u \in H^1(\Omega)$$

i $\gamma u = u_{|\partial\Omega}$ dla wszystkich $u \in C(\overline{\Omega}) \cap H^1(\Omega)$.

[a] Brzeg Ω jest Lipschitzowski (odpowiedniej gładkości), jeśli dla każdego punktu $x \in \partial\Omega$ istnieje otoczenie $\partial\Omega$ tego punktu, które może być reprezentowane jako wykres funkcji Lipschitzowkiej (odpowiednio gładkiej).

Funkcję γu nazywamy śladem u na brzegu $\partial\Omega$.

Kolejnym ważnym twierdzeniem jest tzw. twierdzenie Sobolewa o włożeniu. Tutaj przedstawimy tylko szczególny przypadek potrzebny w przedstawionych dowodach.

Twierdzenie 16.3 (Twierdzenie Sobolewa o włożeniu (ang. *Sobolev embedding theorem*))**.** *Rozpatrzmy Ω ograniczony obszar o brzegu Lipschizowskim w $\mathbb{R}^d$ dla $d = 1, 2, 3$, wtedy - jeśli $2 * k > d$ - istnieje ciągłe włożenie $H^2(\Omega)$ w przestrzeń $C(\overline{\Omega})$ tzn.*

$$\begin{aligned} & & H^2(\Omega) &\subset C(\overline{\Omega}), \\ \exists C > 0 \quad \forall u \in H^k(\Omega) & & \|u\|_{C(\overline{\Omega})} &\leqslant C\|u\|_{H^k(\Omega)}. \end{aligned}$$

Stała $C > 0$ zależy od obszaru Ω.

16.2. Zgodna metoda elementu skończonego

W tym rozdziale przedstawimy ogólne zasady konstrukcji zgodnej metody elementu skończonego. Zgodna metoda oznacza, że przestrzenie elementu skończonego V^h zawarte są w przestrzeni wyjściowej V; w tym przypadku w odpowiedniej przestrzeni Sobolewa.

16.2.1. Element skończony - ujęcie formalne

Najpierw wprowadzimy definicję elementu skończonego za [7], por. także [4] i [2].

Definicja 16.3. — Dla $\tau \subset \mathbb{R}^d$ wielościanu w $\mathbb{R}^d$. (Części brzegu τ leżą na hiperpłaszczyznach i są nazywane ścianami)
— $P_\tau \subset C(\tau)$ jest przestrzenią funkcji wymiaru k określonych na τ (przestrzeń tzw. funkcji kształtu) (ang. *shape functions*)
— $N = (N_1, \ldots, N_k)$ jest baza P_τ^* przestrzeni dualnej do P_τ. (Zbiór stopni swobody elementu). Zazwyczaj te funkcjonały wymagają obliczenia wartości funkcji lub jej pochodnych w punktach, dlatego nazywamy je uogólnionymi warunkami interpolacyjnymi.
wtedy elementem skończonym nazywamy trójkę (τ, P_τ, N).

Definicja 16.4. Dla elementu skończonego (τ, P_τ, N) bazą nodalną tego elementu nazywamy bazę sprzężoną w P_τ do bazy N, tzn. taki układ funkcji z P_τ: $(\phi_1, \ldots, \phi_k)$, że $N_j(\phi_j) = 1$ i $N_j(\phi_l) = 0$ dla $l \neq j$.

Jeśli założymy, że funkcjonały z N są określone i ograniczone na większej lub innej przestrzeni liniowej V, to definiujemy:

Definicja 16.5. Dla elementu skończonego (τ, P_τ, N) definiujemy operator interpolacji $\pi_\tau : V + P_\tau \to P_\tau$:

$$\pi_\tau(f) := \sum_{j=1}^{k} N(f)\phi_j \qquad \forall f \in V$$

dla $(\phi_j)_{j=0}^k$ bazy nodalnej tego elementu.

Jeśli rozpatrujemy podział obszaru na elementy (triangulacje) i każdy element τ jest elementem skończonym, tzn. rozpatrujemy trójkę (τ, P_τ, N_τ), to możemy zdefiniować przestrzeń dyskretną dla danego podziału - zwaną dalej przestrzenią elementu skończonego.

Definicja 16.6. Przestrzenią elementu skończonego V^h dla triangulacji $T_h(\Omega)$ nazywamy dowolną przestrzeń funkcji określonych na Ω takich, że dla funkcji $u \in V^h$ obciętej do elementu $\tau \in T_h$ zachodzi własność

$$u_{|\tau} \in P_\tau.$$

Oczywiście w praktyce elementy skończone są tego samego typu. Często dokładamy na przestrzenie elementu skończonego warunki ciągłości lub dodatkowe warunki na brzegu obszaru.

Definicja 16.3 elementu skończonego dotyczy pojedynczego elementu, a analiza metody elementu skończonego będzie polegała na tym, że wyniki otrzymane na elemencie wzorcowym przenoszą się na dowolny element, o ile wszystkie elementy są skonstruowane przy pomocy przekształceń afinicznych.

Definicja 16.7. Rodzina przestrzeni elementu skończonego V^h dla rodziny triangulacji $T_h(\Omega)$ z $\Omega \subset \mathbb{R}^d$ jest **rodziną afiniczną** pod warunkiem, że istnieje element skończony $(\hat{\tau}, \hat{P}, \hat{N})$ - zwany dalej elementem wzorcowym, i spełnione są następujące warunki: dla dowolnego $\tau_j \in T_h$, istnieje przekształcenie afiniczne $F_j : \hat{\tau} \to \tau$ takie, że dla dowolnej funkcji $u \in V^h$ istnieje $p \in \hat{P}$ takie, że

$$u(x) = p(F_j^{-1}x)$$

oraz dla dowolnego $N_j \in N$ istnieje $\hat{N}_j \in \hat{N}$ takie, że

$$N_j(u) = \hat{N}_j(u \circ F_j).$$

Widzimy, że przekształcenie afiniczne spełnia:

$$F_j\hat{x} = A_j\hat{x} + y_j \quad \hat{x} \in \hat{\tau},$$

dla A_j macierzy nieosobliwej $d \times d$ i y_j ustalonego wektora.

Stwierdzenie 16.2. *Rozpatrzmy afiniczną rodzinę przestrzeni elementu skończonego $\{V^h\}$ dla triangulacji $\{T_h\}$. Wtedy istnieją takie stałe C_1, C_2, że dla elementu triangulacji $\tau_j \in T_h$ i dowolnej funkcji $v \in H^m(\tau_j)$ otrzymujemy:*

$$|\hat{v}|_{H^m(\hat{\tau})} \leqslant C_1\|A_j\|^m|det(A_j)|^{-1/2}|v|_{H^m(\tau_j)},$$
$$|v|_{H^m(\tau_j)} \leqslant C_2\|A_j^{-1}\|^m|det(A_j)|^{1/2}|\hat{v}|_{H^m(\hat{\tau})},$$

gdzie $\hat{v}(\hat{x}) = v(F_j\hat{x})$ dla $\hat{x} \in \hat{\tau}$.

Dowód. Z gęstości funkcji gładkich w H^m możemy założyć, że $v \in C^\infty(\overline{\tau})$. Dowód następnie wynika ze wzoru na różniczkowanie funkcji złożonych:

$$\|\partial^\alpha\hat{v}\|_{L^2(\hat{\tau})} \leqslant C\|A_j\|^m \sum_{|\beta|=m} \|(\partial^\beta v) \circ F_j\|_{L^2(\hat{\tau})}.$$

dla $|\alpha| = m$. Z twierdzenia o podstawianiu otrzymujemy:

$$\|\partial^\alpha\hat{v}\|_{L^2(\hat{\tau})} \leqslant C\|A_j\|^m|det(A_j)|^{-1/2} \sum_{|\beta|=m} \|(\partial^\beta v)\|_{L^2(\tau)}.$$

Sumowanie po wszystkich multiindeksach α o długości m kończy dowód. □

Stwierdzenie 16.3. *Rozpatrzmy afiniczną rodzinę przestrzeni elementu skończonego $\{V^h\}$ dla triangulacji $\{T_h\}$. Wtedy dla $\tau_j \in T_h$ zachodzi:*

$$\|A_j\| \leqslant \frac{\operatorname{diam}(\tau_j)}{\hat{\rho}}, \qquad \|A_j^{-1}\| \leqslant \frac{\operatorname{diam}(\hat{\tau})}{\rho_{\tau_j}},$$

gdzie $\hat{\rho}$ jest średnicą okręgu wpisanego we wzorcowy element $\hat{\tau}$, a ρ_{τ_j} jest średnicą okręgu wpisanego w element τ_j.

Dowód. Widzimy, że

$$\|A_j\| = \sup_{\|z\|=1} \|A_jz\| = \hat{\rho}^{-1} \sup_{\|z\|=\hat{\rho}} \|A_jz\|.$$

Dla dowolnego z o normie $\hat{\rho}$ istnieją $\hat{x}, \hat{y} \in \overline{\hat{\tau}}$, takie, że $z = \hat{x} - \hat{y}$. Zatem biorąc $x = F_j\hat{x}, y = F_j\hat{y} \in \overline{\tau_j}$ otrzymujemy $x - y = F_j(\hat{x}) - F_j(\hat{y}) = A_j(\hat{x} - \hat{y}) = A_jz$, a stąd

$$\|A_j\| \leqslant \hat{\rho}^{-1}\|x - y\| \leqslant \frac{\operatorname{diam}(\tau_j)}{\hat{\rho}}.$$

Drugą nierówność dowodzimy analogicznie. □

Jako wniosek otrzymujemy:

Wniosek 16.1. *Rozpatrzmy regularną rodzinę triangulacji $\{T_h\}$ ze względu na kształt i afiniczną rodzinę przestrzeni elementu skończonego $\{V^h\}$ dla tych triangulacji. Wtedy istnieją takie stałe C_1, C_2, że dla elementu $\tau_j \in T_h$ i dowolnej funkcji $v \in H^m(\tau_j)$ zachodzi*

$$|\hat{v}|_{H^m(\hat{\tau})} \leqslant C(\operatorname{diam}(\tau_j))^m|det(A_j)|^{-1/2}|v|_{H^m(\tau_j)},$$
$$|v|_{H^m(\tau_j)} \leqslant C\rho_{\tau_j}^{-1}|det(A_j)|^{1/2}|\hat{v}|_{H^m(\hat{\tau})},$$

gdzie $\hat{v}(\hat{x}) = v(F_j\hat{x})$ dla $\hat{x} \in \hat{\tau}$.

16.3. Elementy aproksymacji w przestrzeniach Sobolewa H^k

Kolejne twierdzenie pozwala oszacować normę H^m przez półnormę:

Twierdzenie 16.4 (Lemat Deny-Lionsa)**.** *Niech $\tau \subset \mathbb{R}^d$ będzie elementem triangulacji i $l \geqslant 0$. Wtedy istnieje stała $C = C(l, d, \tau)$ taka, że*

$$\inf_{p \in P_l} \|v + p\|_{H^{l+1}(\tau)} \leqslant C|v|_{H^{l+1}(\tau)}.$$

Dowód korzysta z tak zwanej metody zwartości. Można go znaleźć np. w [7], lub [26].

Twierdzenie 16.5. *Rozpatrzmy regularną rodzinę triangulacji $\{T_h\}$ ze względu na kształt i afiniczną rodzinę przestrzeni elementu skończonego $\{V^h\}$ dla tych triangulacji. Jeśli warunki interpolacyjne dla elementu wzorcowego $\hat{N}$ są funkcjonałami liniowymi ograniczonymi na przestrzeni $H^{l+1}(\hat{\tau})$ oraz $P_l \subset \hat{P} \subset H^{l+1}(\hat{\tau})$ dla $0 \leqslant l$, to operator interpolacji nodalnej π_{τ_j} (por. definicję 16.5) jest poprawnie zdefiniowany oraz dla $0 \leqslant m \leqslant l+1$ zachodzi:*

$$|u - \pi_{\tau_j} u|_{H^m(\tau_j)} \leqslant C h_{\tau_j}^{l+1-m} |u|_{H^{l+1}(\tau_j)} \qquad \forall u \in H^{l+1}(\tau_j),$$

dla $h_{\tau_j} = \operatorname{diam}(\tau_j)$ i C zależy od m, l oraz elementu skończonego wzorcowego, i stałej w założeniu regularności ze względu na kształt.

Dowód. Zauważmy, że $\hat{w}(\hat{x}) = w(F_j\hat{x}) = \pi_{\hat{\tau}}\hat{u}(\hat{x})$ dla $\hat{x} \in \hat{\tau}$ i $w = \pi_{\tau_j} u$, co wynika z afiniczności rodziny przestrzeni V^h (por. definicję 16.7).

Stąd na mocy wniosku 16.1 otrzymujemy, że

$$|u - \pi_\tau u|_{H^m(\tau)} = \rho_{\tau_j}^{-m} |det(A_j)|^{1/2} |\hat{u} - \pi_{\hat{\tau}}\hat{u}|_{H^m(\hat{\tau})} \leqslant C\rho_{\tau_j}^{-m} |det(A_j)|^{1/2} (|\hat{u}|_{H^m(\hat{\tau})} + |\pi_{\hat{\tau}}\hat{u}|_{H^m(\hat{\tau})}).$$

Z założeń twierdzenia otrzymujemy teraz:

$$\begin{aligned} |\pi_{\hat{\tau}}\hat{u}|_{H^m(\hat{\tau})} &\leqslant \sum_{j=1}^{k} |N_j(\hat{u})| |\hat{\phi}_j|_{H^m(\hat{\tau})} \leqslant \sum_{j=1}^{k} \|N_j\|_{(H^{l+1}(\hat{\tau}))^*} \|\hat{u}\|_{H^{l+1}(\hat{\tau})} |\hat{\phi}_j|_{H^m(\hat{\tau})} \\ &\leqslant C\|\hat{u}\|_{H^{l+1}(\hat{\tau})} \end{aligned}$$

Oczywiście $\pi_{\hat{\tau}} p = p$ dla dowolnego $p \in \hat{P}$, w szczególności dla p wielomianu z P_l.
Zatem

$$|u - \pi_\tau u|_{H^m(\tau)} \leqslant C\rho_{\tau_j}^{-m} |det(A_j)|^{1/2} \|\hat{u} + p\|_{H^{l+1}(\hat{\tau})} \qquad \forall p \in P_l.$$

Stąd na mocy twierdzenia 16.4 otrzymujemy

$$|u - \pi_\tau u|_{H^m(\tau)} \leqslant C\rho_{\tau_j}^{-m} |det(A_j)|^{1/2} |\hat{u}|_{H^{l+1}(\hat{\tau})}.$$

Z kolei z wniosku 16.1 otrzymujemy

$$|u - \pi_\tau u|_{H^m(\tau)} \leqslant C h_{\tau_j}^{l+1} \rho_{\tau_j}^{-m} |u|_{H^{l+1}(\tau_j)} \leqslant C h_{\tau_j}^{l+1-m} |u|_{H^{l+1}(\tau_j)}.$$

□

Literatura

[1] Uri M. Ascher, Linda R. Petzold. *Computer methods for ordinary differential equations and differential-algebraic equations*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1998.

[2] Dietrich Braess. *Finite elements*. Cambridge University Press, Cambridge, wydanie third, 2007. Theory, fast solvers, and applications in elasticity theory, przetłumaczone z niemieckiego przez Larry L. Schumakera.

[3] K. E. Brenan, S. L. Campbell, L. R. Petzold. *Numerical solution of initial-value problems in differential-algebraic equations*, wolumen 14 serii *Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1996. Przejrzany i poprawiony reprint wydania z roku 1989.

[4] Susanne C. Brenner, L. Ridgway Scott. *The mathematical theory of finite element methods*, wolumen 15 serii *Texts in Applied Mathematics*. Springer, New York, wydanie third, 2008.

[5] J. C. Butcher. *Numerical methods for ordinary differential equations*. John Wiley & Sons Ltd., Chichester, wydanie second, 2008.

[6] P. G. Ciarlet, J.-L. Lions, redaktorzy. *Handbook of numerical analysis. Vol. II*. Handbook of Numerical Analysis, II. North-Holland, Amsterdam, 1991. Finite element methods. Part 1.

[7] Philippe G. Ciarlet. *The finite element method for elliptic problems*, wolumen 40 serii *Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2002. Reprint wydania z roku 1978 [North-Holland, Amsterdam].

[8] Timothy A. Davis. *Direct methods for sparse linear systems*, wolumen 2 serii *Fundamentals of Algorithms*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2006.

[9] James W. Demmel. *Applied numerical linear algebra*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1997.

[10] Maksymilian Dryja, Janina Jankowska, Michał Jankowski. *Metody numeryczne*, wolumen 2. Wydawnictwo Naukowo-Techniczne (WNT), Warszawa, 1982.

[11] Lawrence C. Evans. *Równania Różniczkowe Cząstkowe*. Wydawnictwo Naukowe PWN, Warszawa, 2002. Z języka angielskiego przełożyli Piotr Rybka i Paweł Strzelecki.

[12] E. Hairer, S. P. Nørsett, G. Wanner. *Solving ordinary differential equations. I*, wolumen 8 serii *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, wydanie second, 1993. Nonstiff problems.

[13] E. Hairer, G. Wanner. *Solving ordinary differential equations. II*, wolumen 14 serii *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, wydanie second, 1996. Stiff and differential-algebraic problems.

[14] Janina Jankowska, Michał Jankowski. *Metody numeryczne*, wolumen 1. Wydawnictwo Naukowo-Techniczne (WNT), Warszawa, 1981.

[15] Claes Johnson. *Numerical solution of partial differential equations by the finite element method*. Cambridge University Press, Cambridge, 1987.

[16] Claes Johnson. *Numerical solution of partial differential equations by the finite element method*. Dover Publications Inc., Mineola, NY, 2009. Reprint wydania z roku 1987.

[17] David Kincaid, Ward Cheney. *Numerical analysis. Mathematics of scientific computing*. Brooks/Cole Publishing Co., Pacific Grove, CA, wydanie second, 1996.

[18] David Kincaid, Ward Cheney. *Analiza numeryczna*. Wydawnictwo Naukowo-Techniczne (WNT), 2006.

[19] Andrzej Krupowicz. *Metody numeryczne zagadnień początkowych równań różniczkowych zwyczajnych*. Państwowe Wydawnictwo Naukowe (PWN), Warszawa, 1986.

[20] Randall J. LeVeque. *Finite difference methods for ordinary and partial differential equations*. So-

ciety for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2007. Steady-state and time-dependent problems.

[21] Hanna Marcinkowska. *Dystrybucje, przestrzenie Sobolewa, równania różniczkowe*. Państwowe Wydawnictwo Naukowe (PWN), Warszawa, 1993.

[22] Krzysztof Moszyński. *Rozwiązywanie równań różniczkowych zwyczajnych na maszynach cyfrowych*. Wydawnictwo Naukowo-Techniczne (WNT), Warszawa, 1971.

[23] Andrzej Palczewski. *Równania Różniczkowe zwyczajne. Teoria i metody numeryczne z wykorzystaniem komputerowego systemu obliczeń symbolicznych*. Wydawnictwo Naukowo-Techniczne (WNT), Warszawa, 1999.

[24] William H. Press, Saul A. Teukolsky, William T. Vetterling, Brian P. Flannery. *Numerical recipes*. Cambridge University Press, Cambridge, wydanie third, 2007. The art of scientific computing.

[25] Alfio Quarteroni, Riccardo Sacco, Fausto Saleri. *Numerical mathematics*, wolumen 37 serii *Texts in Applied Mathematics*. Springer-Verlag, New York, 2000.

[26] Alfio Quarteroni, Alberto Valli. *Numerical approximation of partial differential equations*, wolumen 23 serii *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, 1994.

[27] John C. Strikwerda. *Finite difference schemes and partial differential equations*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, wydanie second, 2004.