Więcej o metodach rozwiązywania równań różniczkowych cząstkowych

W. Hebisch.

25 maja 2022

1 Równania w postaci dywergencyjnej

Dla dodatnio określonego operatora A rozwiązywanie równania równania $Ax = b$ jest równoważne minimalizacji formy

$$\langle x, \frac{1}{2}Ax + b\rangle$$

(w praktycznych sytuacjach często pierwotnym problemem jest minimalizacja wyżej, a rozwiązanie równania środkiem do celu). Jeśli x wyżej przebiega przestrzeń nieskończenie wymiarową V to minimalizacja wygląda na trudny problem. Praktyczną metodą przybliżonego rozwiązania problemu minimalizacji jest użycie skończenie wymiarowej podprzestrzeni $W \subset V$. Jeśli przestrzeń W jest dobrze dobrana to rozwiązanie w przestrzeni W będzie bliskie rozwiązaniu w V. Dla W problem sprowadza się do wyznaczenia odpowiedniego operatora skończenie wymiarowego i algebry liniowej (być może stosując metody które omawialiśmy dla równań siatkowych).

Nieco ogólniejszy operator może wyglądać następująco

$$\frac{1}{2}Ax = F(x)$$

gdzie F jest funkcją nielinową. Konkreteniej, operator może wyglądać następująco:

$$Af = (-1)^k \sum_{|\alpha|,|\beta|<k} \partial^\alpha (a_{\alpha,\beta}\partial^\beta f)$$

gdzie $a_{\alpha,\beta}$ zależą od współrzędnej przestrzennej i dla każdego x tworzą macierz dodatnio określoną. Jeśli $a_{\alpha,\beta}$ zależą od f, to wtedy operator będzie nieliniowy. Ogólniej, można rozpatrywać operatory postaci

$$Af = (-1)^k \sum_{|\alpha|\le k} \partial^\alpha u_\alpha(f)$$

gdzie u_α zależny od f i pochodnych f do rzędu k. Przy odpowienich założenich nasz operator będzie eliptyczny. Np. dla A wyżej z $a_{\alpha,\beta}$ jednostanie oszacowanym z dołu zakładając że w A występują tyko człony maksymalnego rzędu mamy

$$\langle Af, f\rangle = (-1)^k \sum_{|\alpha|,|\beta|=k} \langle \partial^\alpha (a_{\alpha,\beta}\partial^\beta f), f\rangle = \sum_{|\alpha|,|\beta|=k} \langle a_{\alpha,\beta}\partial^\beta f, \partial^\alpha f\rangle$$

(ten rachunek wymaga odpowiednich warunków brzegowych, pozwających na całkowanie przez części). Dalej jeśli $a_{\alpha,\beta}$ jest jednostajnie oszacowane z dołu (jako macierz dodatnio określona) to

$$\langle Af, f\rangle \geq c \sum_{|\alpha|=k} \|\partial^\alpha f\|^2_{L^2} \approx \|f\|^2_{H(k)}$$

Jeśli $a_{\alpha,\beta}$ są ograniczone z góry, to wtedy

$$|\langle Af, g\rangle| \leq C\|f\|_{H(k)}\|g\|_{H(k)}$$

A więc zamiast równania

$$Af = b$$

można rozpatrywać równanie

$$\langle Af, g\rangle = \langle b, g\rangle.$$

Powiemy że $f \in H(k)$ jest słabym rozwiązaniem równania $Af = b$ wtedy i tylko wtedy gdy równanie wyżej jest być spełnione dla dowolnego $g \in H(k)$.

Komentarz: Powyżej można rozpatrywać A jako operator z $H(k)$ w $H(-k)$. Wtedy $\langle\cdot,\cdot\rangle$ reprezentuje naturalną dualność między przestrzeniami $H(k)$ i $H(-k)$. Dokładniej, dla dowolnego k funkcje gładkie o nośniku zwartym tworzą podzbiór gęsty w $H(k)$. Na funkcjach gładkich o nośniku zwartym można rozpatrywać zwykły iloczyn skalarny na L^2. Ten iloczyn skalarny rozszerza się do ciągłej formy liniowej względem pierwszego argumentu i antyliniowej względem drugiego argumentu na $H(k) \times H(-k)$. Ta forma zadaje dualność między $H(k)$ a $H(-k)$, tzn. każdy funkcjonał linowy ciągły ϕ na $H(k)$ jest postaci

$$\phi(u) = \langle u, v\rangle$$

dla odpowiedniego $v \in H(-k)$. Ponadto $\|\phi\| = \|v\|$. Dla $k > 0$ przestrzenie$H(-k)$ można by definiować jako to dystrybucje które zadają ciągłe funkcjonały na $H(k)$ i definiować normę na $H(-k)$ jako normę funkcjonału na $H(k)$. Dualność wyżej jest bardzo pożyteczna bo wzory używające dualność $\langle\cdot,\cdot\rangle$ są znacznie prostsze niż wzory używające iloczyn skalarny na $H(k)$.

Przy odpowiednich (dość słabych) założeniach można pokazać że słabe rozwiązania istnieją i są jednoznaczne. Podamy tu jedynie niektóre z prostszych wyników. Powiemy że (ogólnie nieliniowy) operator T z przestrzeni Hilberta V do V' (przestrzeni dualnej) jest monotoniczny jeśli dla każdego $u, v \in V$ mamy

$$\Re(\langle T(u) - T(v), u - v\rangle) \geq 0.$$

Jesli dodatkowo istnieje $m > 0$ taki że dla każdego $u, v \in V$ zachodzi

$$\Re(\langle T(u) - T(v), u - v\rangle) \geq m\|u - v\|^2$$

to powiemy że T jest silnie monotoniczny.

Komentarz: Przy utożsamieniu $V = V'$ operator dodatnio określony jest monotoniczny. Jednakże, są też nieliniowe przykłady. W przypadku rzeczywistym jeśli a jest monotoniczną funkcją jednej zmiennej, to operator

$$T(u)(x) = a(u(x))$$

jest operatorem monotonicznym. Podobnie

$$T(u)(x) = \partial_{x_1}(a(\partial_{x_1} u(x)))$$

jest operatorem monotonicznym.

Powiemy że T jest Lipschitzowski na V ze stałą M jeśli dla każdego $u, v \in V$ mamy

$$\|T(u) - T(v)\| \leq M\|u - v\|.$$

Lemat 1.1 *Jeśli T jest Lipschitzowski i silnie monotoniczny z przestrzeni Hilberta V do V' to dla każdego $f \in V'$ równanie*

$$Tu = f$$

ma jednoznaczne rozwiązanie

Dowód. Przestrzeń Hilberta jest kanonicznie izomorficzna ze swoją przestrznią dualną, więc można przyjąć $H = H'$. Rozważmy operator A_ε zadany wzorem

$$A_\varepsilon(u) = u - \varepsilon(T(u) - f)$$

Mamy

$$\|A_\varepsilon(u) - A_\varepsilon(v)\|^2 = \|(u - v) - \varepsilon(T(u) - T(v))\|^2$$
$$= \|u - v\|^2 - 2\varepsilon\Re(\langle T(u) - T(v), u - v\rangle) + \varepsilon^2\|T(u) - T(v)\|^2$$
$$\leq \|u - v\|^2 - 2\varepsilon m\|u - v\|^2 + \varepsilon^2 M\|u - v\|^2 = (1 - 2\varepsilon m + \varepsilon^2 M)\|u - v\|^2.$$

A więc dla dostatecznie małych $\varepsilon > 0$ operator A_ε jest zwężający i dlatego ma dokładnie jeden punkt stały. Ale punkty stałe A_ε to rozwiązania $Tu = f$. □

Lemat 1.2 *(Laxa-Milgrama) Jeśli $A(u, v)$ jest formą liniową w pierwszym argumencie, antyliniową w drugim argumencie (tzn. $A(u, \lambda v) = \bar{\lambda}A(u, v)$), ograniczoną i koercywną tzn.*

$$\Re(A(u, u)) \geq m\|u\|^2$$

na przestrzeni Hilberta V, to dla każdego ciągłego antyliniowego F na V istnieje jednoznacznie wyznaczone $u \in V$ takie że dla każdego $v \in V$ mamy

$$A(u, v) = F(v).$$

Dowód. Przy ustalonym $u \in V$ forma $\phi_u(v) = A(u,v)$ jest ciągłą formą antyliniową na V, więc istnieje jednoznacznie wyznaczone $T(u) \in V$ takie że

$$A(u,v) = \langle T(u), v\rangle.$$

T jest operatorem liniowym i ciągłym (na mocy ograniczoności A). Na mocy koercywności A operator T jest monotoniczny. A więc do T stosuje się poprzedni lemat. Jeśli $f \in V$ jest taki że $\langle f, v\rangle = F(v)$ to u z poprzedniego lematu daje rozwiązanie. □

2 Metoda elementów skończonych

Metoda elementów skończonych polega na szukaniu przybliżonych rozwiązań w przestrzeni funkcji sklejanych. Dokładniej, dla konkretności rozważamy

$$T(u) = \sum_{|\alpha|\leq k} \partial^\alpha a_\alpha(u).$$

gdzie a_α zależą od współrzędnej przestrzennej x i od pochodnych u do rzędu k włącznie. Przy naturalnym założeniu

$$\|a_\alpha(u)\|_{L^2} \leq C_\alpha \|u\|_{H(k)}$$

operator T odwzorowuje $H(k)$ w $H(-k)$ (pochodne rzędu mniejszego i równego α dają operatory ograniczone z L^2 w $H(-k)$). Naszą przestrzenią Hilberta jest $H(k)$. Zakładamy że T jest operatorem monotonicznym. Dla problemu

$$T(u) = f$$

rozpatrujemy podprzestrzeń $W \subset V$. Rzutowanie P_W z V' na W' definujemu jako obcięcie funkcjonału do przestrzeni W. Niech $T_W = P_W T$. Jeśli T jest monotoniczny lub ściśle monotoniczny to T_W ma tą samą własność, bo dla $u, v \in W$ mamy

$$\langle T_W(u) - T_W(v), u - v\rangle = \langle P_W(T(u) - T(v)), u - v\rangle = \langle T(u) - T(v), u - v\rangle$$

gdzie P_W możemy pominąć bo $u - v \in W$. Podobnie, jeśli T jest Lipschitzowski ze stałą M to T_W jest Lipschitzowski ze stałą nie przekraczającą M. A więc lemat o istnieniu i jednoznaczności działa dla T_W, czyli istnieje jednoznaczne u_W takie że

$$T_W(u_W) = P_W f.$$

Chcielibyśmy oszacować błąd spowodowany przez użycie u_W. Zauważmy najpierw że jeśli u jest dokładnym rozwiązaniem to dla każdego $v \in W$ mamy

$$\langle T(u) - T(u_W), v\rangle = 0.$$

Mianowicie, dla każdego $v \in V$ mamy

$$\langle T(u), v\rangle = \langle f, v\rangle$$

dla każdego $v \in W$ mamy

$$\langle T(u), v\rangle = \langle f, v\rangle$$

i naszą równość otrzymujemy odejmując dwie równości wyżej.

Jeśli T jest operatorem dodatnio określonym oznacza to że u_W jest najlepszym przybliżeniem do u w normie wyznaczonej przez T. Tzn. definiując

$$\langle u, v\rangle_T = \langle Tu, v\rangle$$

dostajemy nowy iloczyn skalarny na V i w tym iloczynie skalarnym u_W jest rzutem ortogonalnym u na W. Czyli

$$\|u - u_W\|_T = \min_{v\in W} \|u - v\|_T.$$

Ogólniej mamy

Lemat 2.1

$$\|u - u_W\| \leq \frac{M}{m} \min_{v\in W} \|u - v\|$$

gdzie M jest stałą Lipschitza dla T zaś m jest stałą koercywności.

Komentarz: Lemat używa normę na V która zwykle jest różna niż $\|\cdot\|_T$.

Dowód: Jak zauważyliśmy wyżej, dla $v \in W$ mamy

$$\langle T(u) - T(u_W), u_W - v\rangle = 0.$$

Teraz

$$m\|u - u_W\|^2 \leq \langle T(u) - T(u_W), u - u_W\rangle = \langle T(u) - T(u_W), u - v\rangle$$

$$\leq \|T(u) - T(u_W)\|\|u - v\| \leq M\|u - u_W\|\|u - v\|.$$

A więc

$$\|u - u_W\| \leq \frac{M}{m}\|u - v\|.$$

Biorąc minimum prawej strony dostajemy wynik. □

Teraz naturalne jest pytanie jak małe może być $\|u - v\|$? Podstawowym parametrem jest wielkość podziału, dalej przez h będziemy oznaczać maksimum średnic komórek podziału. Chcielibyśmy otrzymać oszacowanie typu

$$\min_{v\in W} \|u - v\|_{H(l)} \leq C_{k,l} h^{k-l} \|u\|_{H(k)}. \tag{1}$$

Wymaga to użycia wielominów dostatecznie wysokiego stopnia: łatwo zobaczyć że potrzebne są wielomiany stopnia co najmniej $k - 1$. Są też subtelności

związane z warunkami brzegowymi i przynależnością funkcji sklejanych do właściwego $H(l)$. Zauważmy że jeśli funkcja sklejana jest ciągła to ma jedną słabą pochodną która jest lokalnie ograniczona, a więc w obszarze ograniczonym należy ona do $H(1)$. Jeśli funkcja sklejana ma skok na ściance wymiaru $n-1$, to taka funkcja nie należy do $H(1)$. Ogólniej, jeśli funkcja sklejana jest C^k to należy do $H(k+1)$, zaś nieciągłość k-tej pochodnej na ściance wymiaru $n-1$ wyklucza przynależność do $H(k+1)$. Dla równań rzędu 2 wystarcza nam $H(1)$, czyli ciągłe funkcje sklejane, dla równań rzędu 4 potrzebujemy $H(2)$, czyli funkcje sklejane klasy C^1, ogólniej C^{k-1} dla równań rzędu $2k$.

Jeśli brzeg jest krzywoliniowy, a brzegi naszych komórek są prostoliniowe (zawarte w hiperpłaszczyźnie kowymiaru 1) to odległość brzegu obszaru do brzegu najbliższej komórki może być rzędu h^2. Przy zerowych warunkach brzegowych na brzegu obszaru i na brzegowych komórkach podziału może to dać błąd maksymalmy (L^∞) rzędu h^2. Błąd w przestrzeni $H(k)$ będzie nieco mniejszy, ale dla $l=0$ (czy też $l=1$) i $k=3$ błąd przy brzegu spowoduje że oszacowania (1) nie da się osiągnąć (dla $k=2$ błąd przy brzegu nie wyklucza (1)). Możliwe rozwiązania:

- użyć $k=2$
- zagęścić podział przy brzegu
- użyć komórki z brzegiem krzywoliniowym
- użyć zamianę zmiennych by wyprostować brzeg

Warto tu wspomnieć że w przypadku gdy $l>1$ i brzeg sumy komórek nie pokrywa się z brzegiem obszaru to normę $H(l)$ w (1) liczymy po sumie komórek (na większym zbiorze możemy mieć nieciągłość pierwszej pochodnej u_W).

Nasze podstawowe oszcowanie dla operatora rzędu $2l$ było w normie $H(l)$. A więc zakładając że zachodzą Lemat 2.1, oszacowanie (1) i że dokładne rozwiązanie jest w $H(k)$ mamy

$$\|u-u_W\|_{H(l)} = O(h^{k-l})$$

Dla liniowego dodatnio określonego T, przy założeniu że norma $H(2)$ rozwiązania szacuje się przez wielokrotność normy L^2 prawej strony można pokazać że norma L^2 błędu jest istotnie mniejsza. Mianowicie, niech w będzie rozwiązaniem równania z prawą stroną $u-u_W$

$$Tw = u-u_W.$$

Z założenia

$$\|w\|_{H(2)} \le C_1\|u-u_W\|_{L^2}.$$

Mamy

$$\begin{aligned}\|u-u_W\|_{L^2}^2 &= \langle u-u_W, u-u_W\rangle = \langle u-u_W, Tw\rangle \\ &= \langle T(u)-T(u_W), w\rangle.\end{aligned}$$

Na mocy własności rozwiązań słabych, dla dowolnego $v \in W$ mamy

$$\langle T(u) - T(u_W), v\rangle = 0.$$

A więc

$$\begin{aligned}\|u - u_W\|_{L^2}^2 &= \langle T(u) - T(u_W), w - v\rangle \\ &\leq \|T(u) - T(u_W)\|_{H(-1)}\|w - v\|_{H_1} \\ &\leq M\|u - u_W\|_{H(1)}C_2 h\|w\|_{H(2)} \\ &\leq MC_3 h^{k-1}\|u\|_{H(k)}C_2 hC_1\|u - u_W\|_{L^2} \\ &= C_4 h^k\|u\|_{H(k)}\|u - u_W\|_{L^2}.\end{aligned}$$

Czyli

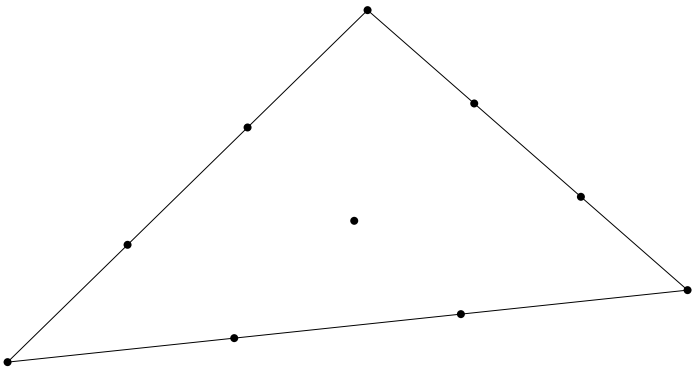
$$\|u - u_W\|_{L^2} \leq C_4 h^k\|u\|_{H(k)}.$$

3 O implementacji metody elementów skończonych

Będziemy zakładać że $\Omega = \sum_{j=1}^{k} \Omega_j$ gdzie Ω_j są prostymi zbiorami takimi jak odcinki, trójkąty (sympleksy w wyższym wymiarze), albo kwadraty (kostki w wyższym wymiarze). W praktyce, by obsłużyć zakrzywiony brzeg należałoby wziąść obraz powyższego zbioru przez odpowiednie odwzorowanie, dla uproszczenia notacji pominiemy to. Zakładamy przekrój $\Omega_j \cap \Omega_k$ albo jest pusty, albo jest wspólną scianką niższego wymiary (czyli wnętrza Ω_j są rozłączne).

3.1 Węzły i funkcjonały węzłowe

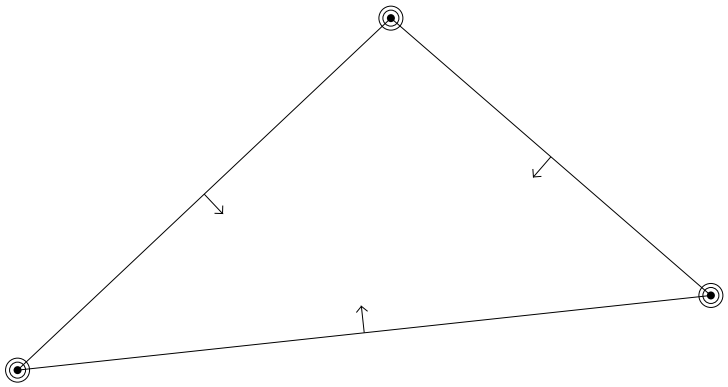
Potrzebujemy w miarę wygodną bazę dla przestrzeni funkcji sklejanych. Robimy to reprezentując wielomiany przez wartości ich i pochodnych w punktach. Dokładniej, zakładamy też że zadane są węzły $x_1, \ldots, x_m \in \Omega$. Wielomian na Ω_j chcemy reprezentować przez wartości i pochodne w punktach x_k takich że $x_k \in \Omega_j$. W najprostszym przypadku mamy reprezentację przez wartości w punktach. Jak już powiedzieliśmy, wielomian liniowy jest jednoznacznie zadany przez wartości w wierzchołkach trójkąta (ogólniej przez wartości w wierzchołkach sympleksu). A więc jeśli Ω_j tworzą triangulację Ω zaś x_k są wierzchołkami triangulacji i używamy przybliżenie liniowe to element W jest jednoznacznie wyznaczony przez wartości w punktach x_k. Podobnie, wielomian kwartatowy jest jednoznacznie wyznaczony przez wartości w wierzchołkach i środkach boków. Podobna reprezenatcja używająca większej ilości punktów działa dla wielomianów wyższych stopni i w wyższych wymiarach. Wartość w punkcie jest funkcjonałem liniowym na W, jeśli używamy reprezentację przez wartości w punktach to odpowiednie funkcjonały nazywamy funkcjonałami węzłowymi.



Na rysunku ilustrujemy rozmieszczenie punktów węzłowych dla wielomianów stopnia 3 na trójkącie.

Można też używać wartości pochodnych jako funkcjonałów węzłowych. Dla wielomianów stopnia 3 na trójkącie można użyć wartości i wartości pierwszych pochodnych w środkach boków i dodatkowo wartość w środku trójkąta. Niestety, odpowiednia funkcja sklejana dalej może mieć nieciągłą pochodną bo pochodne w kieronku normalnym do boków nie muszą się zgadzać. Jednakże, używając pochodne jako funkcjonały węzłowe zmniejszamy wymiar przestrzeni W, co zwykle jest korzystne. Mianowicie, w typowej triangulacji z M wierzchołkami mamy w przybliżeniu $2M$ trójkątów i $3M$ krawędzi. Reprezentacja stopnia 3 przez wartości w punktach ma w przybliżeniu M punktów węzłowych w wierzchołkach, $2 \cdot 3M = 6M$ punktów węzłowych na krawędziach i $2M$ punktów węzłowych w środkach trójkątów, co daje łącznie wymiar $9M$. Reprezentacja przez pochodne w wierzchołkach ma $3M$ funkcjonałów w wierzchołkach (wartość i dwie pochodne na wierzchołek) i w przybliżeniu $2M$ na wartości w środku trójkąta, razem $5M$.

Do uzyskania C^1 potrzebujemy co najmniej stopień 5. Do wyznaczenia wielomianu stopnia 5 wystarcza zadanie pochodnych do rzędu 2 w wierzchołkach i pochodnych normalnych w środkach boków. To gwarantuje zgodność wielomianów na wspólnych bokach i zgodność pochodnych normalnych do boków, czyli ciągłość pochodnych.



Na rysunku pokazujemy odpowiednie punkty węzłowe. Punkty z powójnymi kółkami reprezentują wartości i pochodne do rzędu 2. Strzałki reprezentują pochodne normalne.

Przy użyciu wielomianów stopnia 5 jak wyżej potrzebujemy 6 współczynników na wierzchołek triangulacji i po jednym na środek krawiędzi, czyli w przybliżeniu przy M wierzchołkach potrzebujemy $6M$ współczynników w wierzchołkach i $3M$ na środkach krawędzi, razem $9M$, a więc porównywalnie do reprezentacji przez wartości w punktach dla stopnia 3. Oczywiście, stosunkowo mała ilość współczynników wynika z tego że warunek C^1 wprowadza zależności między współczynnikami na sąsiednich trójkątach.

W przypadku kwadratów (czy kostek) naturalne są elementy produktowe. Na odcinku wielomian stopnia 3 jest wyznaczony przez wartości i pierwsze pochodne na końcach. Odpewiednie wielomiany na kwadracie mają stopień 3 względem kazdej zmeiennej z osobna, ale łączny stopień wynosi 6. Taki wielomian jest wyznaczony przez 4 współczynniki w każdym wierzchołku: wartość, dwie pierwsze pochodne i drugą pochodną mieszaną. Przy podziale liczba kwadratów jest w przybliżeniu równa liczbie wierzchołków, a więc wymiar odpowiedniej przestrzeni na M wierzchołkach to $4M$. Przy tym otrzymujemy funkcje klasy C^1. Podobną konstrukcję można przeprowadzić dla wielomianów stopnia 5 względem każdej zmiennej. Teraz trzeba 9 współczynników w każdym wierzchołku, wymiar przestrzemi w przybliżeniu wynosi więc $9M$ dla M wierzchołków. Otrzymujemy funkcje klasy C^2.

3.2 Macierz układu

W słabym sformułowaniu rozwiązujemy równanie

$$\langle Tu, v\rangle = \langle f, v\rangle.$$

Dla uproszczenia zakładając że operator jest rzędu 2 i liniowy możemy napisać:

$$Tu = \sum_{k,l} \partial_k(a_{k,l}\partial_l u) + \sum_k a_k \partial_k u + a_0 u$$

(teoretycznie mogą też być człony typu $\partial_l(b_l u)$ ale dla nas nie wprowadzają one istotnej zmiany). Dla uproszczenia będziemy też zakładać że nasze przestrzenie Hilberta są rzeczywiste, tak że pominiemy sprzężenia zespolone. Całkując przez części dostaniemy

$$\langle Tu, v\rangle = -\sum_{k,l}\langle a_{k,l}\partial_l u, \partial_k v\rangle + \sum_k \langle a_k \partial_k u, v\rangle + \langle a_0 u, v\rangle$$

Zakładając jak poprzednio że $\Omega = \sum_j \Omega_j$, że wnętrza Ω_j są rozłączne zaś brzegi mają miarę 0 możemy napisać

$$\langle a_{k,l}\partial_l u, \partial_k v\rangle = \int_\Omega a_{k,l}(x)\partial_l u(x)\partial_k v(x)dx = \sum_j \int_{\Omega_j} a_{k,l}(x)\partial_l u(x)\partial_k v(x)dx.$$

Zauważmy że dla $u, v \in W$ pojedyncze wyrażenie

$$\int_{\Omega_j} a_{k,l}(x)\partial_l u(x)\partial_k v(x)dx$$

zwiera całkę z wielomianu $\partial_l u(x)\partial_k v(x)$ razy współczynnik $a_{k,l}$ po prostym zbiorze Ω_j. Dla $a_{k,l}$ które jest stałe czy ogólniej jest wielomianem te całki są łatwe do dokładnego obliczenia. W innych przypadkach można jej przybliżać (np. przybliżając $a_{k,l}$ wielomianem) lub obliczać numerycznie. Jak powiedzieliśmy elementy W reprezentujemy przez wartości funkcjonałów węzłowych ϕ_m. Na Ω_j możemy znaleźć odpowiednią bazę dualną tak że

$$v|_{\Omega_j} = \sum \phi_m(v)\eta_{m,j}.$$

Wyżej dla uproszczenia notacji używamy pełny zestaw funkcjonałów węzłowych, jeśli $\phi_m(v)$ jest niepotrzebne na Ω_j to po prostu przyjmujemy że $\eta_{m,j} = 0$. Do bazy zaliczamy tylko niezerowe $\eta_{m,j}$.

Piszemy teraz

$$A_{j,k,l,m,n} = \int_{\Omega_j} a_{k,l}(x)\partial_l \eta_{m,j}\partial_k \eta_{n,j}dx$$

Podobnie piszemy

$$A_{j,k,m,n} = \int_{\Omega_j} a_k(x)\partial_l \eta_{m,j}\eta_{n,j}dx,$$

$$A_{j,m,n} = \int_{\Omega_j} a_0(x)\eta_{m,j}\eta_{n,j}dx,$$

$$B_{j,m,n} = -\sum_{k,l} A_{j,k,l,m,n} + \sum_k A_{j,k,m,n} + A_{j,m,n}$$

i

$$B_{m,n} = \sum_j B_{j,m,n}.$$

Macierz B wyżej nazywamy macierzą układu (w angielskiej literaturze jest też nazwa *stiffness matrix*). Skoro

$$u|_{\Omega_j} = \sum \phi_m \eta_{m,j}$$

to

$$-\sum_{k,l} \int_{\Omega_j} a_{k,l}(x)\partial_l u(x)\partial_k v(x)dx + \sum_k \int_{\Omega_j} a_k(x)\partial_k u(x)v(x)dx$$

$$+ \int_{\Omega_j} a_0 u(x)v(x)dx = \sum_{m,n} B_{j,m,n}\phi_m(u)\phi_n(v)$$

A więc

$$\langle Tu, v\rangle = \sum_{m,n} B_{m,n}\phi_m(u)\phi_n(v)$$

czyli macierz B faktycznie reprezentuje nasz operator gdy reprezentujemy funkcje przez wartości funkcjonałów węzłowych. W praktyce liczba komórk podziału jest duża w stosunku do wymiaru używanej przez nas przestrzeni wielomianów na Ω_j. Powoduje to że macierz B jest macierzą rzadką, choć proporcja niezerowych elementów jest większa niż w macierzach które pojawiają się przy prostych przybliżeniach różnicowych.

Przy programowaniu metody elementów skończonych trzeba wybrać reprezentację macierzy B. Nie zawsze warto jawnie tworzyć macierz B. Typowe algorytmy iteracyjne potrzebują tylko obliczać produkty Bu dla $u \in W$. Lecz do tego wystarczają $B_{j,m,n}$. Zauważmy że przy ustalonym j niezerowe elementy $B_{j,m,n}$ tworzą stosunkowo małą macierz B_j. Wartość Bu można obliczać jako

$$Bu = \sum_j B_j u.$$

Oczywiście łączna liczba niezerowych elementów macierzy B_j jest większa lub równa ilości niezerowych elementów macierzy B. Lecz dla macierzy rzadkich trzeba też pamiętać położenie niezerowych elementów. Dla macierzy B_j wystarcza odpowiedniość między współrzędymi u a współrzędnymi na Ω_j. Jeśli Ω_j jest trójkątem i używamy wielomiany stopnia 3 to wymiar przerzeni wielomianów to 10. Czyli dla ustalonego j wystarcza 10 liczb by opisać związek elementów B_j (którą można reprezentować w sposób gęsty) z współrzędnymi u. Dla macierzy B takich liczb trzeba więcej. Dokładniej, dla pojedynczego trójkąta macierz B_j ma $10 \cdot 10 = 100$ elementów. Macierz B ma w przybliżeniu 76 elementów na trójkąt czyli tworząc B ze 100 elementów oszczędzimy 24. Lecz przy prostej reprezentacji każdy element B wymaga informację o położeniu, tzn. dodatkową liczbę całkowitą. $76 - 10 = 66$ liczb całkowitych zwykle zajmuje więcej miejsca niż 24 liczby zmiennopozycyjne, więc nie tworząc macierzy

B oszczędzimy pamięć. Przy mnożeniu, mnożąc przez B_j musimy wykonać 10 indeksowanych dostępów do pamięci by wybrać potrzebne współrzędne, samo mnożenie przez B_j możemy wykonać w miarę szybką procedurą dla pełnych macierzy, po czym użyć 10 indesowancych dostępów by uaktualnić składowe wyniku. Przy najprostszej rzadkiej reprezentacji B każde mnożenie przez element B prowadzi do indeksowanego dostępu do pamięci. A więc koszt iloczynu Bu przy rzadkiej reprezentacji B może być wyższy od kosztu mnożenia przy niejawnej reprezentacji za pomocą B_j.

Uwaga: Używając macierz B musimy opowiednio reprezentować f. Chodzi o to że gdy używamy funkcjonały węzłowe jako współrzędne to iloczyn skalarny we współrzędnych zwykle nie pokrywa się z iloczynem skalarnym w L^2. Dostaniemy macierze

$$C_{j,m,n} = \int_{\Omega_j} \eta_{m,j}\eta_{n,j}$$

i

$$C_{m,n} = \sum_j C_{j,m,n}.$$

Wtedy

$$\langle f, v\rangle = \sum_{n,m} C_{m,n}\phi_m(f)\phi_n(v) = \sum_n f_n\phi_n(v)$$

gdzie

$$f_n = \sum_m C_{m,n}\phi_m(f).$$

Przykład: Dla $T = -\partial_x^2$ na odcinku i funkcji sklejanych stopnia 2 reprezentowanych przez wartości w punktach macierz B_j na odcinku jednostkowym to

$$\frac{1}{3}\begin{bmatrix} 7 & -8 & 1 \\ -8 & 16 & -8 \\ 1 & -8 & 7 \end{bmatrix}.$$

Macierz C_j to

$$\frac{1}{15}\begin{bmatrix} 2 & 1 & -1 \\ 1 & 8 & 1 \\ -1 & 1 & 2 \end{bmatrix}.$$

3.3 Rozwiązywanie macierzy układu

Metody rozwiązywania macierzy układu są podobne do metod używanych dla przybliżeń różnicowych. W najprostszym przypadku równanie $Bu = f$ można rozwiązywać bezpośrednio korzystając z algorytmów dla macierzy rzadkich. Niekiedy, szczególnie dla małych lecz nieregularnych problemów może to być preferowana metoda. Można stosować metody iteracyjne (np. z przyspieszaniem Czebyszewa czy metodę sprzężonych gradientów). Można też stosować metody wielosiatkowe. Ze względu na to że w metodzie elementów skończonych

macierze nie są tak rzadkie jak przy przybliżeniach różnicowych, metody bezpośrednie mogą być przydatne dla nieco większych problemów. Warto też zauważyć że przy metodach bezpośrednich główny koszt to rozkład LU, co oznacza że wielokrotne rozwiązywanie układu z różnymi prawymi stronami jest niewiele bardziej kosztowne od rozwiązania pojedynczego układu. To zjawisko oznacza że może być atrakcyjna metoda dwu siatek: na mniejszej stosujemy metodą bezpośrednią, na większej iterację, której duży krok najpierw rozwiązuje problem na mniejszej siatce, potem ograniczona liczba iteracji na większej siatce poprawia rozwiązanie. Jest to uproszczona wersja metody wielosiatkowej. W porównaniu do metody bezpośredniej rozkład LU na mniejszej siatce jest dużo tańszy niż na pełnej siatce. Dzięki temu że rozkład LU obliczamy tylko raz użycie metody bezpośredniej na mniejszej siatce może być tańsze od wielokrotnego stosowania metody iteracyjnej (co ma miejce w pełnej metodzie wielosiatkowej).

3.4 Modyfikacje

Dotychczas zakładaliśmy że warunki brzegowe są spełnione dokładnie i że $W \subset H(k)$. Jednakże, zapisując formę dwuliniową odpowiadającą naszemu operatorowi jako

$$\langle Tu, v\rangle \sum_j \int_{\Omega_j} \sum_{|\alpha|\leq k} a_\alpha(u)\partial^\alpha v$$

widzimy że ma ona naturalne rozszerzenia na funkcje sklajane u, nawet jeśli $u \notin H(k)$. Jeśli $u \notin H(k)$ to całkowanie przez części wprowadza wyrazy brzegowe na brzegach Ω_j i można by się obawiać znacznej utraty dokładności. Okazuje się że wyrazy brzegowe nie muszą być dokładnie 0, wystarczy że są małe. Wyrazy brzegowe będą małe gdy wartości u i pochodnych będą się zgadzać w niektórych punktach brzegu.

Dlaczego takie podejście się przydaje? Daje ono większą swobodę by spełnić dodatkowe warunki jak np. zasady zachowania. Łatwiej buduje się przestrzenie dla operatorów wyższego rzędu. Łatwiej też uwzględnia się zakrzywiony brzeg.

Istotną zaletą metody elementów skończonych jest możliwość rozwiązywania operatorów z nieciągłymi współczynnikami. Dokładniej, interesują nas operatory ze współczynnikami mającymi skoki. Np. mamy płytę składającą się z częsci wykonanych z różnych meteriałów które mają różne parametry. W takim wypadku oczekujemy że skok współczynnika oznacza mniejszą regularność rozwiązania, np. skok pochodnej. Naturalne podejście dobiera Ω_j tak by skoki współczynników były na brzegach Ω_j. Przy skoku współczynnika nie nakładamy warunku ciągłości na pochodne (choć może się pojawić warunek zgodności różny od ciągłości pochodnej).

Jeśli oczekujemy że rozwiązanie jest w niektórych miejscach mniej regularne to ma sens zastosowanie gęstszego podziału w tym miejscu. Jeśli mamy szacowanie błędu to można adaptacyjnie (automatycznie) zagęszczać podział.

Wcześniej głównym sposobem osiągnięcia zbieżności było zagęszczanie siatki (podziału). Jednakże można też podwyższać stopień wielomianu. W niskich wymiarach często daje to większą dokładność niż zagęszczanie podziału. Miano-

wicie, jeśli dokładne rozwiązanie jest bardzo regularne to podwyższanie stopnia może dać wykładnicze malenie błędu, czyli metodę nieskończonego rzędu.

Warunek dodatniej określoności czy monotoniczności może być ograniczający. Jednym ze sposobów by go uzyskać jest zastąpienie operatora liniowego A przez A^*A. Oznacza to podwojenie rzędu operatora, co zwykle jest niekorzystne. Lecz dla operatora A rzędu 1 dostajemy A^*A rzędu 2 co często daje dobry efekt. Przy tym czasami operatory wyższego rzędu daje się zastąpić przez systemy rzędu 1.