Równania cząstkowe, metody na siatkach

W. Hebisch.

27 kwietnia 2022

1 Przykładowe równania

Niech

$$\Delta f = \sum_{j=1}^{n} \partial_{x_j}^2 f$$

będzie operatorem Laplace'a na $\mathbb{R}^n$. Równanie Laplace'a to

$$\Delta f = 0. \tag{1}$$

To równanie ma bardzo wiele rozwiązań, aby uzyskać jednoznaczne rozwiąznie potrzebujemy dodatkowe warunki (warunki brzegowe). Niech $\Omega \subset \mathbb{R}^n$ będzie zbiorem otwartym ze zwartym domknięciem i regularnym brzegiem. W zagadnieniu Dirichleta przy zadanym g na $\partial\Omega$ (brzegu Ω) szukamy rozwiązania równania (1) w Ω takiego że

$$f = g$$

na $\partial\Omega$ (aby to równanie miało sens potrzebna jest wystarczająca regularność g). Przykładowe problemy fizyczne które prowadzą do zagadnienie Dirichleta w wymiarze $n = 3$ to wyznaczanie temperatury w obszarze przy zadanej temperaturze na brzegu czy wyznaczanie potencjału elektrycznego w obszarze przy zadanym potencjale na brzegu.

W zagadnieniu Neumanna zadajemy pochodną f w kierunku normalnym do brzegu (to wymaga silniejszych założeń).

Ogólniej, można zadać pochodną w wybranym kierunku czy relację między funkcją a pochodną.

Równanie Poissona jest prostą modyfikacją (1):

$$\Delta f = g$$

która jednak wpływa istotnie na rozwiązanie. Podobnie jak dla równania Laplace'a trzeba wprowadzić warunki brzegowe. Przykładowy problem fizyczny to wyznaczanie pola elektrycznego znając rozkład ładunku (czy Newtonowskiego potencjału grawitacyjnego przy zadanym rozkładzie masy).

Powyższe równania opisywały sytuacje statyczne, ale interesujące są też równania opisujące ewolucję w czasie. Naprostszym z nich jest równanie ciepła

$$\partial_t u(x,t) = \Delta_x u(x,t) + f(x,t)$$

gdzie Δ_x oznacza że Laplasian liczymy tylko względem zmiennych x. Powyżej u jest szukaną funkcją zaś f jest zadane. Podobnie jak dla równania Laplace'a czy Poissona potrzebne są warunki brzegowe. Ponadto potrzebny jest warunek początkowy, tzn. zadajemy u dla pewnego t_0, np $t_0 = 0$ i szukamy rozwiązania dla $t > t_0$. Dla równania ciepła naturalne są podobne warunki brzegowe jak dla równania Laplace'a. Dodatkowo dla równania ciepła (czy Poissona) można rozważać zerowy warunek brzegowy w nieskończoności tzn. rozważać rozwiązania dążące do 0 gdy x dąży do nieskończoności (dla równania Laplace'a dałoby to rozwiązanie zerowe). Można też badać okresowy warunek brzegowy, tzn. brać pod uwagę tylko rozwiązania okresowe względem zmiennych x.

Przykładowy problem fizyczny to wyznaczanie temperatury u przy zadanej temperaturze początkowej, temperaturze na brzegu (warunek brzegowy Dirichleta) i dopływie ciepła (za to odpowiada f). Alternatywnie można zadać przepływ ciepła przez brzeg, prowadzi to do warunku brzegowego Neumanna. Inny problem to dyfuzja, wtedy u to gęstość cząstek, f to tworzenie cząstek we wnętrzu obszaru, warunek początkowy daje początkową gęstość, warunek brzegowy Dirichleta zadaje gęstość na brzegu, warunek Neumanna przepływ przez brzeg.

Wariantem równania ciepła jest niestacjonarne równanie Schrödingera

$$\partial_t u(x,t) = i\Delta_x u(x,t) - iV(x)u(x,t)$$

gdzie i jest jednostką urojną zaś V jest zadaną funkcją rzeczywistą (potencjałem). Opisuje ono ewolucję funkcji falowej z potencjałem V. Pomnożenie Δ przez i bardzo mocno zmienia charakter rozwiązań. Równanie ciepła daje się rozwiązać wstecz tylko dla niezwykle regularnych danych i rozwiązanie wstecz jest niestabilne przy typowych normach. Równanie Schrödingera jest symetryczne względem czasu, można je rozwiązywać zarówno w przód jak i wstecz. Naturalny w zagadniech fizycznych jest zerowy warunek brzegowy w nieskończoności, dokładniej, zwykle interesują nas rozwiązania całkowalne w kwadratem względem x przy ustalonym t.

Wspomnijmy też równanie fali

$$\partial_t^2 u(x,t) = \Delta_x u(x,t)$$

Opisuje ono rozchodzenie się fal, np. mechanicznych czy elektromagnetycznych.

W równaniach wyżej część najwyższego rzędu miała stałe współczynniki. W fizycznych problemach odpowiada to założeniu że materiał jest jednorodny. Ogólniej zamiast operatora Laplace'a można rozważać operator typu

$$\sum_{j=1}^{n} \partial_{x_j}(h\partial_{x_j})$$

gdzie h jest funkcją opisującą własności materiału (np. przewodnictwo cieplne, czy przenikalność elektryczną). Ogólniej, niektóre materiały są anizotropowe i ich własności trzeba opisywać macierzą. Prowadzi to do operatora

$$\sum_{j=1}^{n}\sum_{k=1}^{n} \partial_{x_j}(h_{j,k}\partial_{x_k})$$

Dla liniowych równań ewolucyjnych takich jak równanie ciepła, niestacjonarne równanie Schrödingera czy równanie falowe często stosuje się metodę rozwinięcia na funkcje własne. Mianowicie rozpatrujemy równanie

$$L(u) = \lambda u$$

gdzie L jest jednym z operatorów wyżej (np. $L = \Delta$) zaś λ jest liczbą nazywaną wartością własną. u wyżej nazywamy funkcją własną (lub wektorem własnym). Jak poprzednio, potrzebne są warunki brzegowe. W najprostszych przykładach funkcje własne daje się wyznaczyć jawnymi wyrażeniami (lub są one dobrze zbadane), ale ogólnie do ich wyznaczenia potrzebne są metody numeryczne. W przypadku stacjonarnego operatora Schrödingera

$$L(u) = -\Delta u + Vu$$

wartości własne odpowiadają energii, zaś stan stabilny (podstawowy) odpowiada najmniejszej wartości własnej. Toteż znajdowanie najmniejszej wartości własnej operatora Schrödingera ma duże znaczenie w fizyce i chemii. Podobnie, wartości własne operatora Laplace'a odpowiadają asymptotycznemu zachowaniu równania ciepła, pozwalają też wyznaczyć prawdopodobieństowo że w wyniku dyfuzji cząstka opuści obszar.

Nasze równania wyżej były liniowe. Jednakże, w praktyce trzeba też uwzględniać efekty nieliniowe. Np. przenikalność magnetyczna i elektryczna substancji jest mocno nieliniowa przy dużych zmianach pól. Prowadzi to do operatorów typu

$$N(u) = \sum_{j=1}^{n} \partial_{x_j}(h(u)\partial_{x_j})u$$

czy

$$N(u) = \sum_{j=1}^{n} \partial_{x_j}(h(|\nabla u|^2)\partial_{x_j})u$$

gdzie

$$|\nabla u|^2 = \sum_{j=1}^{n} |\partial_{x_j} u|^2.$$

W szczególności, problem wyznaczania powierzchni o minimalnym polu rozpiętej na danej krzywej czyli zagadnienie minimalnej powierzchni ma ostatnią postać wyżej z

$$h(|\nabla u|^2) = \frac{1}{\sqrt{1 + |\nabla u|^2}}.$$

Efekty nieliniowe pojawiają się też przy ewolucji, np. nieliniowe dyfuzje, nieliniowe równanie ciepła czy nieliniowe równania falowe.

W bardziej skomplikowanych sytuacjach pojawiają się układy równań. Np. dyfuzja cząstek naładowanych prowadzi do układu równań wiążącego gęstość cząstek i pole. Pola elektryczne i magnetyczne zmieniają się w czasie zgodnie z

układem równań Maxwella. Przepływ cieczy nieściśliwych prowadzi do układu równań Nawiera-Stokesa.

Nasze przykłady były równaniami rzędu 2, bo takimi będziemy się głównie zajmować. Jednakże w praktycznych zagadnieniach mogą się pojawić równania rzędu 1 czy też wyższego rzędu.

2 Teoria równań cząstkowych

Gdy szukamy rozwiązania dobrzy by wiedzieć czy rozwiązanie istnieje i czy jest jednoznaczne. Dla równań zwyczajnych twierdzenie Picarda podaje prosty warunek który wystarcza w większości zastosowań. Dla równań cząstkowych sytuacja jest znacznie bardziej skomplikowana. Dla niektórych równań nie ma rozwiązań lokalnie, tzn. nawet jak pominiemy warunki brzegowe. Równanie ciepła z rozsądnymi warunkami brzegowymi i początkowymi ma regularne rozwiązanie w przód, ale zwykle nie daje się go rozwiązać wstecz. Warunki brzegowe wymagają uwagi, jeśli je rozpatrujemy punktowo to czasami nie mają sensu. Rozwiązania równania zwyczajnego zwykle są bardziej regularne niż równanie. Dla równań cząstkowych rozwiązania mogą być nieregularne. Dla wielu typów równań istnienie, regularność czy jednoznaczność rozwiązań są otwartymi problemami badawczymi.

Warto tu też wspomnieć że o ile równania zwyczajne można badać używając zwykłą definicję pochodnej i supremum wartości bezwględnej (tzn. normę L^∞)) do pomiaru wielkości funkcji, to dla równań cząstkowch potrzebne są pochodne uogólnione (słabe), zaś najlepiej zachowującą się normą jest norma L^2. Przy tym w konkretych zagadnienich często istotną częścią teorii jest dobranie odpowiedniej normy, tak by równanie dobrze się zachowywało względem wybranej normy. Słabe rozwiązania i normy są związane z metodą elementów skończonych o której będziemy później mówić.

Ze względu na komplikacje ogólnej teorii wspomnę tu tylko o trzech klasycznych klasach równań, dla których sporo wiadomo.

Jedna klasa równań to równania eliptyczne, przykładem jest równanie Poissona. Dla takich równań rozwiązania są bardziej regularne niż równanie. Dla regularnych równań liniowych wystarcza norma L^2 i np. równanie typu

$$L(u) = g$$

z regularnym eliptycznym L rzedu m ma rozwiązanie u bardziej regularne niż g. Dokładniej, jeśli g ma s pochodnych w L^2, to u ma $m+s$ pochodnych w L^2. Równania zwyczajne bez osobliwości (spełniające założenia twierdzenia Picarda) są eliptyczne w tym sensie. Dla równań nieliniowych potrzebne są też normy L^p i normy Hölderowskie. Jednakże dalej daje się pokazać że równanie poprawia regularność. Co do istnienia rozwiązań, to przy rozsądnych założeniach dość łatwo pokazać że równania eliptyczne mają słabe rozwiązania (które potencjalnie są niezbyt regularne) i że takie rozwiązanie jest jednoznaczne. Dalej teoria regularności pozwala pokazać że faktycznie te słabe rozwiązania zachowują się

lepiej. Nasze podstawowe metody będą nastawione na równania eliptyczne a podstawowym przykładem będzie równanie Poissona.

Drugą klasą równań są równania paraboliczne. Przykładem tu jest równanie ciepła. Tu nawet dla niezbyt regularnych warunków początkowych (powiedzmy dla $t_0 = 0$) rozwiązanie staje się bardziej regularne. Dla równań parabolicznych jest analogiczna teoria jak dla równań eliptycznych: pokazuje się istnienie i jednoznaczność słabych rozwiązań i twierdzenia że słabe rozwiązanie faktycznie jest bardziej regularne. Numerycznie nasze podejście jest związane z potraktowaniem równania parabolicznego jako równania zwyczajnego o wartościach wektorowych z przestrzeni nieskończnie wymiarowej, gdzie działanie na przestrzeni nieskończnie wymiarowej jest przez operator eliptyczny.

Trzecią klasą równań są równia hiperboliczne jak równanie fali czy niestacjonarne równanie Schrödingera. Tym razem nie ma efektu regularyzacji rozwiązania, ale dalej daje się pokazać istnienie i jednoznaczność rozwiązań. Znowu specjalną rolę odgrywa współrzędna czasowa, a na współrzędnych przestrzennych mamy operator eliptyczny.

Można tu też wsponieć że pojawiają się zdegenerowane wersje równań wyżej, np. aby nieliniowa wersja Laplasianu była eliptyczna to funkcja h powinna być dodatnia. Jeśli w pewnych punktach h się zeruje to mamy operator zdegenerowany, dalej daje się pokazać istnienie rozwiązań ale degeneracja zwykle wiąże się z nieregularnością. Jeśli daje się kontrolować nieregularność (np. jest punkt osobliwy czy podzbiór mniejszego wymiaru) to przy pewnej uwadze dalej się stosują metody eliptyczne.

3 Pochodne na siatkach

Przy rozwiązywaniu równań cząstkowych jedną z możliwych reprezentacji jest reprezentacja przez wartości w wybranych punktach i przybliżanie pochodnych różnicami skończonymi. Dla uproszczenia naszym wyborem jest regularna siatka prostokątna, w wymiarze dwa $x_{i,j} = x_{0,0} + h(i,j)$. W praktyce bardzo ważne są równania rzędu 2, w szczególności równania z operatorem takim jak laplasian. Dla pochodnych rzędu 2 różnica symetryczna daje przybliżenie rzędu 2:

$$\partial_x^2 f(x) = \frac{f(x-h) + f(x+h) - 2f(x)}{h^2} + O(h^2).$$

Stosując ten wzór po współrzędnych dostaniemy przybliżenie do laplasianu. Podobny wzór pozwala przybliżać pochodne mieszane. Mianowicie, niech

$$\Gamma_{h,h} f = f(x+h, y+h) + f(x-h, y-h) - f(x+h, y-h) - f(x-h, y+h).$$

Wtedy

$$\partial_{x,y} f(x,y) = \frac{\Gamma_{h,h} f}{2h^2} + O(h^2).$$

Stosując więcej punktów można uzyskać przybliżenia wyższego rzędu.

Biorąc kombinacje liniowe można przybliżać dowolne operatory różniczkowe rzędu 2. Jednakże, zachowanie rozwiązania zależy bardzo od struktury równania. Dla równań hiperbolicznych i parabolicznych dobrze jest dobrać układ współrzędnych tak by równanie opisywało ewolucję w czasie i zastosować specjalne metody. Dla równań eliptycznych typowo można po prostu rozwiązywać otrzymany układ na siatce.

Dodajmy że podane wyżej dyskretne przybliżenie do Laplasianu zachowuje się dość dobrze. Np. spełniona jest wersja zasady maksimum: rozwiązanie osiąga maksimum tylko na brzegu siatki, tzn. punkt siatki taki że jego wszyscy czterej sąsiedzi należą do siatki nie może być punktem maksimum. Dyskretena wersja równania ciepła na prostą interpretacją probabilistyczną: opisuje błądzenie przypadkowe najbliższego sąsiada, tzn. jeśli w czasie t jesteśmy w punkcie p z siatki to w czasie $t+1$ przechodzimy z równym prawdopodobieństwem do jednego z sąsiadów. Warunek brzegowy Dirichleta oznacza że błądzenie kończy się gdy mamy przejść do punktu spoza obszaru. Warunek Neumanna oznacza że odbijamy się od brzegu, tzn. krok prowadzacy poza obszar zastępujemy krokiem w przeciwną stronę.

4 Obszary i warunki brzegowe

Na prostej jedyne zbiory spójne to przedziały (być może niewłaściwe). W wyższych wymiarach mamy znacznie więcej zbiorów spójnych i samo opisanie dziedziny rozwiązania może być skomplikowne. By uniknąć kłopotów zakładamy że szukamy rozwiązania w obszarze z gładkim i zwartym brzegiem. Wtedy dla przybliżeń siatkowych wystarczy wziąść dostatecznie mały krok h i jako przybliżenie obszaru użyć punkty zawarte we wnętrzu obszaru.

Dla równań eliptycznych rzędu 2 postaci

$$L(f) = h \tag{2}$$

gdzie L jest operatorem różniczkowym typowe zagadnienie zawiera warunki brzegowe. Najprostszy jest warunek Dirichleta. Warunek Dirichleta

$$f(x) = g(x)$$

możemy reprezentować w ten sposób że przedłużamy g w gładki sposób na zewnętrze obszaru. Dla niektórych punktów z wnętrza przybliżona różnica będzie zależała od wartości f poza obszarem. Zamiast tych wartości bierzemy wartości dane przez przedłużony warunek brzegowy. Jest to szczególnie proste wtedy gdy warunek brzegowy jest zerem. W efekcie zamiast równania (2) rozpatrujemy równanie

$$Av = w$$

gdzie A jest przbliżem różnicowym do L, zaś w reprezentuje h i człony z L odpowiadające wyrazowm brzegowym.

Przykład: Rozpatrujemy laplasian w obszarze $U = (0,1) \times (0,1)$ z zerowym warunkiem brzegowym, czyli równanie

$$-(\Delta f)(x,y) = h(x,y)$$

dla $(x,y) \in U$ z warunkiem $f(0,y) = f(1,y) = f(x,0) = f(x,1) = 0$. Biorąc krok $h = \frac{1}{5}$ i zaczynając siatkę w początku układu punkty z $i,j \in \{1,2,3,4\}$ leżą we wnętrzu U. Zapisując równania w postaci macierzowej wygodnie jest użyć cztery indeksy. Mamy wtedy

$$(Av)_{(i,j)} = \sum_{(k,l)} a_{(i,j),(k,l)} v_{(k,l)}$$

gdzie $i,j,k,l \in \{1,2,3,4\}$ zaś $a_{(i,j),(k,l)}$ spełnia

$$a_{(i,j),(i,j)} = 4,$$

(czyli elementy diagonalne to 4) i

$$a_{(i,j),(k,l)} = -1$$

o ile (i,j) różni się od (k,l) o wektor jednostkowy. Pozostałe elementy A są zerami. Macierz takiej postaci nazywa się macierzą wstęgową. Wygląda ona tak:

$$\left(\begin{array}{cccc|cccc|cccc|cccc}
4 & -1 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
-1 & 4 & -1 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & -1 & 4 & -1 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & -1 & 4 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\hline
-1 & 0 & 0 & 0 & 4 & -1 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & -1 & 0 & 0 & -1 & 4 & -1 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & -1 & 0 & 0 & -1 & 4 & -1 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & -1 & 0 & 0 & -1 & 4 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 \\
\hline
0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 4 & -1 & 0 & 0 & -1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & -1 & 4 & -1 & 0 & 0 & -1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & -1 & 4 & -1 & 0 & 0 & -1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & -1 & 4 & 0 & 0 & 0 & -1 \\
\hline
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 4 & -1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & -1 & 4 & -1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & -1 & 4 & -1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & -1 & 4
\end{array}\right)$$

Dla obszaru nie będącego prostokątem macierz ma mniej regularną strukturę ale dalej pojawia się podobne zachowanie.

5 Rozwiązywanie równań siatkowych

Dla dwuwymiarowej siatki $N \times N$ mamy macierz kwadratową wymiaru N^2. Rozwiązanie odpowiadającego jej układu równań metodą eliminacji Gaussa wymaga rzędu N^6 operacji i N^4 komórek pamięci. Jednakże macierz równania

siatkowego jest rzadka, większość jej elementów to zera. Toteż naturalne jest szukanie szybszych metod rozwiązania.

Dla macierzy rzadkich często stosuje się specjalne warianty eliminacji Gaussa, wykonując obliczenia prawie tylko na niezerowych wyrazach. Jest tu problem: elementy które początkowo są zerami mogą w trakcie stać się niezerowe (po angielsku mówi się o fill in). W szczególności odwrotność macierzy rzadkiej typowo na wszystkie elementy niezerowe. Dlatego nie oblicza się odwrotności a zadowala redukcją do postaci trójkątnej, która może być rzadka. Odpowiednio wybierając elementy podstawowe można to ograniczyć pojawianie się nowych wartości (fill in), ale nie da się tego całkiem wyeliminować. Organizacji obliczeń na macierzach rzadkich poświęcone są całe książki. Tu nie będziemy patrzeć na detale takich metod. Wspomnimy jedynie że w macierzy wyżej wszystkie niezerowe elementy są w odległości co najwyżej N od diagonali. Dla takich macierzy wystarcza rzędu N^5 operacji i N^3 pamięci.

5.1 Metoda sprzężonych gradientów

Metoda sprzężonych gradientów bazuje na minimalizacji formy kwadratowej

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

Widać że pochodna względem x jest równa 0 wtedy i tylko wtedy gdy $Ax = b$. Dla macierzy dodatnio określonych forma kwadratowa wyżej osiąga minimum dokładnie w punktach takich że $Ax = b$.

Okazuje się że formę kwadratową wyżej można mnimializować w następujący sposób:

1. zaczynamy z dowolnego x_0, bierzemy $r_1 = Ax_0 - b$, $d_1 = -r_1$,
2. dla i zaczynając od 1 iteracyjnie bierzemy $x_i = x_{i-1} + \alpha_i d_i$ gdzie α_i jest dobrane tak by zminimalizować formę kwadratową na linii $x_{i-1} + \alpha d_i$,
3. jeśli i jest równe wymiarowi przestrzeni lub osiągneliśmy potrzebną dokładność to kończymy,
4. bierzemy $r_{i+1} = Ax_i - b$, $d_{i+1} = -r_i + \beta_i d_i$ gdzie β_i jest dobrane tak by d_i i d_{i+1} były ortogonalne,
5. powiększamy i o 1 i przechodzimy do kroku 2.

Kluczowe własności metody wyżej:

- wektory d_i wyżej są ortogonalne,
- x_i minimalizuje formę kwadratową na hiperpłaszczyźnie

$$x_0 + \mathrm{lin}\{d_1, \ldots, d_i\} = x_0 + \mathrm{lin}\{r_1, Ar_1, \ldots, A^{i-1}r_1\}.$$

Obliczeniowo pojedynczą iterację metody sprzężonych gradientów wygodnie zapisać następującymi wzorami:

$$\alpha_i = \frac{\langle r_i, r_i \rangle}{\langle Ad_i, d_i \rangle},$$

$$r_{i+1} = r_i + \alpha_i A d_i,$$

$$\beta_i = \frac{\langle r_{i+1}, Ad_i \rangle}{\langle Ad_i, d_i \rangle},$$

$$d_{i+1} = -r_{i+1} + \beta_i d_i$$

co wymaga jednego mnożenia wektora przez macierz (czyli obliczenie Ad_i), trzy produkty skalarne i dwie kombinacje liniowe wektorów.

A więc dla macierzy symetrycznych i dodatnio określonych $n \times n$ metoda sprzężonych gradientów pozwala obliczyć rozwiązanie kosztem rzędu n mnożeń macierz-wektor i porównywalnej liczby operacji na wektorach. Dla równań siatkowych jak wyżej $n = N^2$ zaś koszt mnożenia wektora przez macierz jest proporcjonaly do n. A więc metodą sprzężonych gradientów można rozwiązać równanie kosztem rzędu N^4 operacji.

5.2 Rozdzielanie zmiennych i FFT

Jeśli operator, obszar i warunki brzegowe mają strukturę produktową, tzn.

$$L(f) = L_x f + L_y f$$

gdzie L_x jest częścią operatora działającą na zmiennej x zaś L_y jest częścią operatora działającą na zmiennej y to funkcje własne i wartości własne mają strukturę produktową. Mianowocie, jeśli ϕ_j jest pełnym układem funkcji własnych dla L_x ψ_k jest pełnym układem funkcji własnych dla L_y to produkty $\phi_j \psi_k$ tworzą pełnym układ własnych dla L. Dokładniej, jeśli

$$L_x \phi_j = \lambda_j \phi_j,$$

$$L_y \psi_k = \eta_k \psi_k,$$

i ϕ_j oraz ψ_k tworzą układy zupełne, to

$$L(\phi_j \psi_k) = (\lambda_j + \eta_k) \phi_j \psi_k$$

i produkty tworzą układ zupełny. Ważnym przykładem operatora o strukturze produktowej jest Laplasian w obszarze będącym produktem i przy warunkach brzegowych Dirichleta (lub Neumanna). Wtedy można dobrać siatkę tak by dyskretny operator też był produktowy. Jeśli mamy szybką metodę przechodzenia od wartości w punktach do rozkładu na funkcje własne i z powrotem to można ją użyć do rozwiązywania równania

$$L(f) = g$$

Mianowicie, jeśli

$$g = \sum_{j=0}^{N} c_j \phi_j$$

to

$$f = \sum_{j=0}^{N} \frac{c_j}{\lambda_j} \phi_j.$$

Dla uproszczenia notacji we wzorze wyżej pominełem strukturę produktową. Dla pochodnej funkcje własne to exponenty, podobnie dla laplasianu. Uwzględniając warunki brzegowe Dirichleta dla jednej zmiennej funkcje własne na odcinku $(0, \pi)$ to $\sin(kx)$ z $k = 1, 2, \ldots$ (odpowiednia wartość własna to k^2). A więc rozwinięcie na funkcje własne sprowadza się do sinusowego przekształcenia Fouriera. W dyskretnej wersji można użyć algorytm FFT który działa w czasie rzędu $O(N \log(N))$. Struktura produktowa pozwala stosować to samo podejście w przypadku większej ilości zmiennych.

5.3 Metody iteracyjne

Innym sposobem jest obliczanie przybliżonego rozwiązania metodą w stylu iteracji punktu stałego. Dokładniej, dla niezerowego γ równanie

$$Ax = b$$

jest równoważne równaniu

$$x = x + \gamma(Ax - b).$$

Jeśli $\|I + \gamma A\|$ jest mniejsza od 1 to iteracja punktu stałego $x_{i+1} = x_i + \gamma(Ax_i - b)$ jest zbieżna do rozwiązania x_∞. Zauważmy że startując z x_0 błąd $x_i - x_\infty$ spełnia zależność

$$x_{i+1} - x_\infty = x_i + \gamma(Ax_i - b) - x_\infty = x_i + \gamma(Ax_i - b) - x_\infty - \gamma(Ax_\infty - b)$$

$$(x_i - x_\infty) + \gamma A(x_i - x_\infty).$$

A więc $x_i - x_\infty = P(A)(x_0 - x_\infty)$ gdzie $P(z) = (1 + \gamma z)^i$. Aby oszacować błąd zakładamy że A jest dodatnio określoną macierzą symetryczną. Możemy wtedy zdiagonalizować A przy pomocy przekształceń ortogonalnych. Innymi słowy istnieje baza ortogonalna e_j taka że $Ae_j = \lambda_j e_j$. Pozwala to oszacować normę $P(A)$ przez działanie na wektorach własnych

$$\|P(A)\| = \max \|P(A)e_j\| = \max |1 + \gamma\lambda_j|^i.$$

Jeśli λ_1 jest najmiejszą wartością własną A zaś λ_n jest największą to łatwo sprawdzić że najmniejszą wartość wyżej dostaniemy biorąc $\gamma = \frac{-2}{\lambda_n + \lambda_1}$ i wtedy

$$\|P(A)\| = \frac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1}.$$

Niestety, dla typowego operatora w stylu laplasianu na dwuwymiarowej siatce mamy $\lambda_1 \approx 1$, $\lambda_n \approx n$ co prowadzi do rzędu $n\log(\varepsilon)$ iteracji by osiągnąc dokładność ε co jest gorsze od metody sprzężonych gradientów.

Jednakże wynik można znacznie poprawić zmieniając γ z kroku na krok. Wtedy jako $P(A)$ możemy otrzymać dowolny wielomian taki że $P(0) = 1$. Aby zminimalizować błąd chcemy zminimalizować supremum P na odcinku $[\lambda_1, \lambda_n]$ pod warunkiem $P(0) = 1$. To zagadnienie rozwiązują przetransformowane wielomiany Czebyszewa T_i, tzn. $P(z) = cT_i(az+b)$ gdzie $az+b$ przesztalca przedział $[\lambda_1, \lambda_n]$ na $[-1, 1]$ zaś c jest tak dobrane by $P(0) = 1$. Okazuje się że wtedy wystarcza i rzędu

$$\sqrt{\frac{\lambda_n}{\lambda_1}}\log(\varepsilon).$$

tzn. $\sqrt{n}\log(\varepsilon)$ iteracji.

Jeśli używamy metodę sprzężonych gradientów jako metodę dokładną, to trzeba n interacji i wtedy powyższe jest znacznie lepsze. Z drugiej strony, po k iteracjach metoda sprzężonych gradientów minimalizuje formę

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

na hiperpłaszczyźnie $x_0 + \mathrm{lin}\{r_1, Ar_1, A^{k-1}r_1\}$. Oznacza to że w odpowiedniej normie dostajemy najlepsze przybliżenie do rozwiązanie w tej hiperpłaszczyźnie. Metoda iteracji typu Czebyszewa z tą samą ilością kroków działa w tej samej hiperpłaszczyźnie.

5.4 Metody wielosiatkowe

Jeśli dla iteracji punktu stałego użyjemy γ o module nieco mniejszym od optymalnego, np. $\gamma = \frac{-3}{2(\lambda_n+\lambda_1)}$ to biorąc $\alpha = \frac{\lambda_n+\lambda_1}{6}$ i $V_n = \mathrm{lin}\{e_j : \lambda_j \geq \alpha\}$ dla $v \in V_n$ mamy

$$\|P(A)v\| \leq \|v\| \max_{j \geq n/2} |(1+\gamma\lambda_j)|^i \leq \left(\frac{3}{4}\right)^i \|v\|.$$

A więc składowa błędu z podprzestrzeni V_n maleje szybko, zaś problem sprawia podprzestrzeń $W_n = \mathrm{lin}\{e_j : \lambda_j < \alpha\}$. Dla naszego przykładowego problemu przestrzeń W_n ma znacznie mniejszy wymiar niż cała przestrzeń i składa się ze stosunkowo regularnych funkcji (mówi się często że V_n reprezentuje składowe wysokiej częstotliwości zaś W_n składowe niskiej częstotliwości). Nasuwa to pomysł by najpierw uzyskać rozwiązanie na mniejszej siatce, np. z krokiem pomnożonym przez 2, przedłużyć to rozwiązanie na oryginalną siatkę i użyć jako przybliżenie początkowe. Liczymy na to że wtedy błąd w przestrzeni W_n będzie mały i szybkość zbieżności będzie taka jak na V_n.

Przy naiwnej realizacji pomysłu wyżej jest kłopot, mianowicie przejście z mniejszej do większej siatki wprowadza pewien błąd, w efekcie błąd w przestrzeni W_n będzie pochodził głównie z przejścia między siatkami i na pierwszy

rzut oka nie da się go zmniejszyć do zera. Jednakże modyfikacja pomysłu wyżej działa, zilustrujemy pomysł dla równania (2) przy założeniu że L jest operatorem liniowym. Metoda działa w wielu etapach (krokach), przy tym duże etapy są podzielone na mniejsze. Jeden duży etap produkuje przybliżone rozwiązanie f_1 równania (2) ze stosunkowo dużym błędem, ale tak by

$$\|L(f_1) - h_1\| \leq \frac{1}{2}\|h_1\|$$

gdzie $h = h_1$ jest prawą stroną równania (2). Następnie bierzemy $h_2 = -(L(f_1) - h_1)$ i szukamy f_2 tak by

$$\|L(f_2) - h_2\| \leq \frac{1}{2}\|h_2\|$$

i podobnie dla $f_3, \ldots, f_i$. Zauważmy że biorąc

$$f = \sum_{k=1}^{i} f_k$$

mamy

$$L(f) - h = L(f_i) - h_i$$

czyli na mocy założeń o f_i mamy

$$\|L(f) - h\| \leq \frac{1}{2^i}\|h\|.$$

A więc po i dużych etapach możemy uzyskać dowolnie dużą dokładność.

Pozostaje wyjaśnić jak zrealizować duży etap. Można to zrobić jak w pomyśle wyżej: najpierw produkujemy rozwiązanie na mniejszej siatce (nie musi być bardzo dokładne), przedłużamy na większą i kilka razy stosujemy iterację punktu stałego na dużej siatce. Istotne przy tym jest to że ilość iteracji na dużej siatce nie zależy od rozmiaru siatki, tak że praca na dużej siatce zależy liniowo od rozmiaru tej siatki. Na mniejszej siatce metodę wyżej stosujemy rekursywnie: znowu wystarcza ustalona liczba iteracji na mniejszej siatce. Liczba iteracji na mniejszej będzie większa od ilości iteracji na dużej siatce, ale można dobrać parametry tak by praca bezpośrednio na mniejszej siatce była mniejsza, ale proporcjonalna do pracy na dużej siatce. W efekcie metoda działa na wielu siatkach, na małych siatkach jest stosunkowo dużo iteracji, ale ponieważ te siatki są mniejsze to łącznie praca na mniejszych siatkach jest ograniczona przez wielokrotność pracy na dużej siatce. Innymi słowy, duży etap wymaga rzedu N^2 operacji, tzn. złożoność jest liniowa ze względu na rozmiar problemu. Typowo zadowalna nas ustalona dokładność zaś np. 30 dużych etapów daje błąd względny rzędu 2^{-30} czyli w praktyce metoda wielosiatkowo wymaga wysiłku liniowo zależnego od rozmiaru problemu, co pozwala rozwiązywać bardzo duże problemy.

Nasze uzasadnienie dla metod wielosiatkowych zakładało że macierz na siatce jest symeryczna. Jednakże bardziej skomplikowana analiza pokazuje że symetria

nie jest krytyczna, podobne oszacowanie można uzyskać dla problemów niesymetrycznych (być może kosztem pewnego zwiększenia ilości iteracji). W praktyce stałe i parametry metody odgrywają istotną rolę i wiele wysiłku wkłada się w uzyskanie lepszych stałych.

5.5 Problemy nieliniowe

Dla równań nieliniowych jedną z możliwości jest użycie metody Newtona. Naiwnie użyta metoda Newtona może mieć problem z powodu niezbyt dobrego przybliżenia początkowego i wymagać rozwiązania wielu problemów liniowych, co dałoby znaczy koszt. Jednakże metodę Newtona daje się połączyć ze schematem wielosiatkowym. W miarę dobre przybliżenie początkowe dostajemy na mniejszej siatce. W początkowych iteracjach nie ma potrzeby rozwiązywania problemów liniowych z dużą dokładnością, co oszczędza czas obliczeń. W efekcie można rozwiązywać równania nieliniowe na dużych siatkach kosztem proporcjonalnym do kosztu rozwiązania problemu liniowego.