WYKŁAD 2: Klasyczne metody iteracyjne dla układów równań liniowych

1. Wstęp

Wiele zagadnień numerycznych prowadzi do układów równań liniowych o dużej liczbie niewiadomych. Typowe zagadnienia z dziedziny mechaniki ośrodków ciągłych (mechanika płynów, dynamika gazów, mechanika odkształcalnych ciał stałych, zagadnienia transportu ciepła i masy, itp.) wymagają rozwiązania (często - wielokrotnego) układów liniowych o rozmiarach od rzędu kilkuset/kilkadziesiąt tysięcy (zagadnienia 2D) do kilkudziesięciu/kilkuset milionów (a nawet więcej!) w złożonych geometrycznie zagadnieniach 3D.

Cechą tych układów liniowych jest to, że ich macierze współczynników zawierają relatywnie mało elementów niezerowych. O takich macierzach mówimy, że są RZADKIE (ang. sparse matrices). W typowych zastosowaniach współczynnik „nadłości" (sparseness factor) - tj. stosunek liczby elementów niezerowych do wszystkich elementów - może być naprawdę bardzo mały, szczególnie dla dużych rozmiarów macierzy. Np. przy zastosowaniu metody różnic skończonych do równania z operatorem Laplace'a w 2D będzie on wynosił $\sim 5/N$ gdzie N to wymiar macierzy (powiedzmy $N \sim 10^4 \div 10^5$).

Do rozwiązywania układów liniowych o takich rozmiarach i strukturze nie stosuje się metod „dokładnych" (czyli nieiteracyjnych) typu eliminacji Gaussa (lub jej pochodnych)

Istnieją zasadniczo dwa podstawowe powody:

1) koszt numeryczny metod dokładnych - nawet właściwie zaimplementowanych pod kątem efektywnego użycia dla macierzy rzadkich - jest gigantyczny, tj. proporcjonalny do kwadratu rozmiaru układu (wsp. proporcjonalności może być również znaczny, a jego dokładna wartość zależy od tzw. rozległości „pasma" macierzy, czyli maksymalnej odległości elementów niezerowych od głównej diagonalnej)
2) podczas procesu rozwiązania potrzeba znacznie więcej pamięci operacyjnej niż do przechowania samych niezerowych elementów macierzy. Powodem tego jest efekt pojawiania się niezerowych wartości w miejscach gdzie wartości elementów macierzy były zerowe (tzw. fill-in). Efekt ten może zwiększyć zapotrzebowanie na pamięć nawet o rząd wielkości!

Reasumując: wielkie układy liniowe z macierzami rzadkimi rozwiązujemy metodami iteracyjnymi!

Podział metod iteracyjnych:

Metody iteracyjne → Klasyczne (Jacobi, Gauss-Seidel, SOR, itd)
Metody iteracyjne → Krylowa (gradienty sprzężone, GMRES i wiele innych)

2. Wybrane informacje nt. stacjonarnych liniowych procesów iteracyjnych.

Stacjonarnym liniowym procesem iteracyjnym nazywamy proces zadany wzorem:

$$x^{n+1} = Gx^{n} + w \qquad (1)$$

x^{n+1} – wektor x w iteracji $n+1$-szej
G – stała, zadana macierz
w – zadany wektor
x^{n} – wektor x w iteracji n-tej

przy czym zadany jest wektor startowy x^0.

Punktem (wektorem) stałym procesu (1) nazywamy wektor x_* taki, że

$$x_* = Gx_* + w \qquad (2)$$

Zauważmy, że $(I-G)x_* = w$. Wyznaczenie punktu stałego iteracji (1) oznacza zatem rozwiązanie układu z macierzą $I-G$ i wektorem prawych stron w (albo zadania równoważnego).

Zobaczmy od czego zależy zbieżność procesu (1). Wprowadźmy wektor błędu:

$$e^n := x^n - x_* \quad ; \quad n = 0, 1, 2, \ldots$$

Zauważmy, że odejmując stronami (1) i (2) otrzymamy

$$x^{n+1} - x_* = G(x^n - x_*) \Rightarrow e^{n+1} = Ge^n \qquad (3)$$

Wektory błędu spełniają zatem naszą formułę iteracyjną z $W = 0$.

Udowodnimy następujące

TWIERDZENIE: Warunkiem koniecznym i wystarczającym zbieżności procesu iteracyjnego

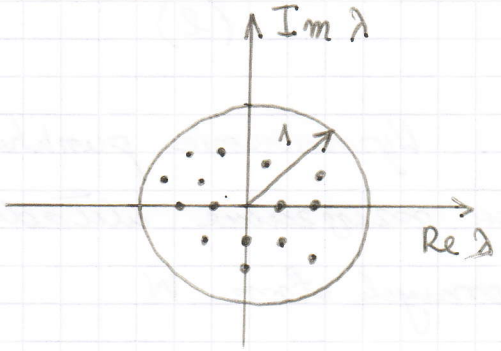
$$x^{n+1} = G x^n + W$$

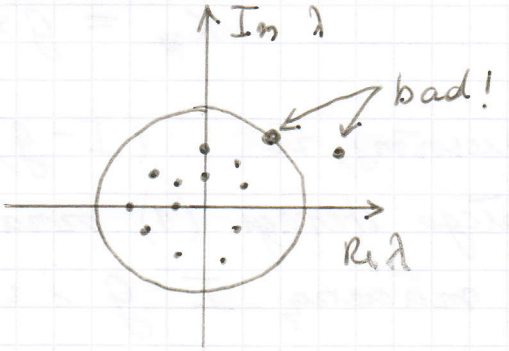
przy dowolnym x^0

do punktu stałego x_* (tj. $x_* = G x_* + W$) jest, aby wszystkie wartości własne macierzy G spełniały warunek

$$|\lambda| < 1. \quad \text{(nierówność ostra!)}$$

Komentarz: innymi słowy wymaga się, aby wszystkie wartości własne G (są to - na ogół - liczby zespolone) znajdowały się we wnętrzu koła o promieniu 1





Wielkość $\varrho(G) = \max |\lambda(G)|$ nazywamy PROMIENIEM SPEKTRALNYM macierzy G. Twierdzenie mówi, że dla zbieżności potrzeba i wystarcza, aby $\varrho(G) < 1$!

Dowód:

Dowiedziemy tylko konieczności warunku. Załóżmy - wbrew założeniu - że istnieje wartość własna $\hat{\lambda}$ macierzy G taka, że $|\hat{\lambda}| \geq 1$. Z teorii macierzy wynika, że istnieje wówczas wektor (własny) $\hat{v}$ taki, że:

$$G\hat{v} = \hat{\lambda}\hat{v} \qquad (4)$$

Przyjmijmy wektor startowy $x^0 = x_* + \hat{v}$. Wówczas $e^0 = \hat{v}$ i

$$e^1 = Ge^0 = G\hat{v} = \hat{\lambda}\hat{v}$$

Dalej $e^2 = Ge^1 = G(\hat{\lambda}\hat{v}) = \hat{\lambda}G\hat{v} = \hat{\lambda}^2\hat{v}$ itd.

Ogólnie $e^n = \hat{\lambda}^n\hat{v} = \hat{\lambda}^n e^0$

a stąd, przy dowolnie dobranej normie wektorowej mamy

$$\|e^n\| = |\hat{\lambda}|^n \cdot \|e^0\|.$$

Ponieważ $|\hat{\lambda}| \geq 1$ więc zachodzi dla dowolnego $n > 0$ nierówność $\|e^n\| \geq \|e^0\|$ co oczywiście zaprzecza możliwości, że $\lim_{n\to\infty} \|e^n\| = 0$.

KONIEC

Wyjaśnimy znaczenie wielkości promienia spektralnego $\varrho(G)$. Dla uproszczenia założymy, że macierz G ma pełny układ wektorów własnych $\{v_1, v_2, \ldots, v_n\}$, $n = \dim G$. Niech ponadto $\{\lambda_1, \lambda_2, \ldots, \lambda_n\}$ będzie układem odpowiadających tym wektorom wartości własnych, tj.

$$Gv_j = \lambda_j v_j \quad , \quad j = 1, 2, \ldots, n$$

Ponieważ $\{v_1, \ldots, v_n\}$ to układ liniowo niezależny to początkowy wektor błędu można zapisać jako ich kombinację liniową

$$e^0 = x^0 - x_* = \alpha_1 v_1 + \alpha_2 v_2 + \ldots + \alpha_n v_n$$

$\alpha_1, \alpha_2, \ldots, \alpha_n$ - jednoznacznie określone liczby.

Wówczas

$$e^1 = \sum_{j=1}^{n} \alpha_j G v_j = \sum_{j=1}^{n} \alpha_j \lambda_j v_j$$

$$e^2 = \sum_{j=1}^{n} \alpha_j \lambda_j G v_j = \sum_{j=1}^{n} \alpha_j \lambda_j^2 v_j \quad \text{itd.}$$

Ogólnie: $$e^n = \sum_{j=1}^{n} \alpha_j \lambda_j^n v_j$$

Dla dużych wartości n otrzymana suma zdominowana będzie przez składnik odpowiadający największej co do modułu wartości własnej (zakładamy, że stojący przy tym składniku współczynnik α nie jest „przypadkiem" zerowy). Oznacza to, że dla dużych n

$$e^n \approx \alpha_{max} \lambda_{max}^n v_{max}$$

$$e^{n+1} \approx \alpha_{max} \lambda_{max}^{n+1} v_{max}$$

Stąd: $$\mu = \frac{\|e^{n+1}\|}{\|e^n\|} \approx \frac{|\alpha_{max}| \rho^{n+1}(G) \|v_{max}\|}{|\alpha_{max}| \rho^n(G) \|v_{max}\|} = \rho(G)$$

gdzie μ nazywamy asymptotycznym tempem zbieżności. Jak widać im mniejsze $\rho(G)$ tym lepiej!

Zabiegi pomocnicze zmierzające do uzyskania jak największej redukcji promienia spektralnego $\rho(G)$ nazywamy poprawianiem uwarunkowania procesu iteracyjnego (ang. preconditioning)

3. Klasyczne metody iteracyjne

a) metoda Jacobiego

$Ax = b$ – układ do rozwiązania.

Zakładamy, że dla dowolnego $j = 1, 2, .., n = \dim A$ elementy diagonalne $a_{jj} \neq 0$. Jeśli tylko macierz A jest nieosobliwa da się to osiągnąć metodą zmiany numeracji równań i/lub niewiadomych.

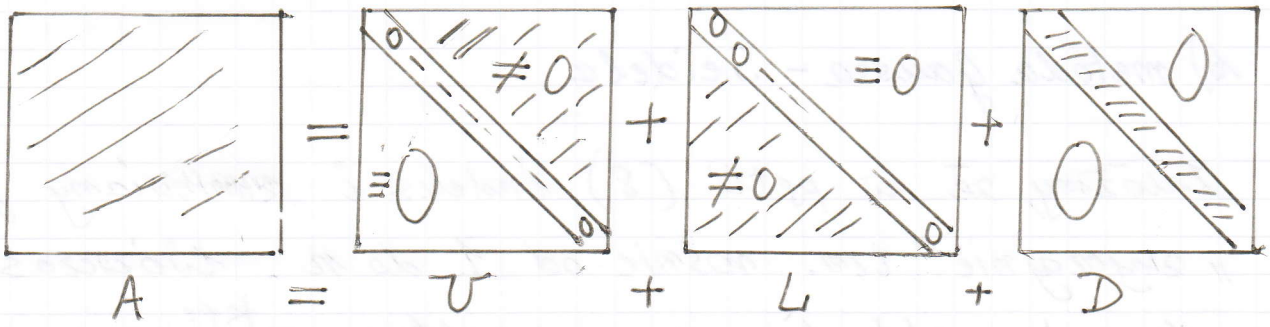
Definiujemy macierze:

$$U: \quad u_{ij} = \begin{cases} a_{ij}, & i < j \\ 0, & i \geq j \end{cases}$$

$$L: \quad l_{ij} = \begin{cases} 0, & i \leq j \\ a_{ij}, & i > j \end{cases}$$

$$D: \quad d_{ij} = \begin{cases} a_{ii}, & i = j \\ 0, & i \neq j \end{cases}$$

Obrazkowo ...



Skoro $A = U + L + D$ to układ przyjmuje postać

$$(U + L + D)x = b$$

$$Dx = -(U + L)x + b$$

$$x = -D^{-1}(U + L)x + D^{-1}b \qquad (5)$$

Otrzymaliśmy układ równań dla wektora (punktu) stałego dla następującego procesu iteracyjnego:

$$x^{k+1} = -D^{-1}(U+L)x^k + D^{-1}b \qquad (6)$$

Jest to właśnie metoda Jacobiego. Mamy dla niej

$$G_J = -D^{-1}(U+L), \quad W_J = D^{-1}b. \qquad (7).$$

Rozpisując (6) na składowe, otrzymamy pętlę do wykonania w każdej iteracji metody Jacobiego

for $i \in \{1, 2, .., n\}$ do:

$$x_i^{k+1} = \frac{1}{a_{ii}}\left(b_i - \sum_{j=1}^{i-1} a_{ij}x_j^k - \sum_{j=i+1}^{n} a_{ij}x_j^k\right) \qquad (8)$$

end;

Zauważmy, że sposób zapisania pętli w (8) ma sugerować, że faktyczna kolejność indeksu i nie ma znaczenia – aktualizacja każdego elementu wektora x^{k+1} jest kompletnie niezależna od pozostałych elementów. Wynika stąd wniosek, że pętlę można wykonać RÓWNOLEGLE!

b) metoda Gaussa - Seidela

Załóżmy, że w pętli (8) indeks i zmieniany jest „zwyczajnie" tzn. rośnie od 1 do n. Wówczas elementy aktualizowanego wektora x^{k+1} o numerach od 1 do i-1 mogą być od razu wykorzystane do obliczenia x_i^{k+1} (bo zostały policzone wcześniej w tej właśnie iteracji)

Otrzymaliśmy proces iteracyjny opisany następująco

$$\text{for } i = 1:n \text{ do}$$
$$x_i^{k+1} = \frac{1}{a_{ii}}\left(b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{k+1} - \sum_{j=i+1}^{n} a_{ij} x_j^{k}\right) \qquad (9)$$
$$\text{end;}$$

Jest to metoda iteracyjna Gaussa-Seidela.
Wyprowadzimy postać macierzowo-wektorową tej metody.
Z formuły (9) wynika, że

$$x^{k+1} = -\mathcal{D}^{-1}\left[L x^{k+1} + U x^{k}\right] + \mathcal{D}^{-1} b$$

czyli

$$Dx^{k+1} = -Lx^{k+1} - Ux^{k} + b$$
$$(\mathcal{D}+L)x^{k+1} = -Ux^{k} + b$$
$$\Downarrow$$
$$x^{k+1} = -(\mathcal{D}+L)^{-1} U x^{k} + (\mathcal{D}+L)^{-1} b$$

Mamy zatem:

$$x^{k+1} = G_{gs} x^{k} + W_{gs} \qquad (10)$$

gdzie

$$\begin{cases} G_{gs} = -(\mathcal{D}+L)^{-1} U \\ W_{gs} = (\mathcal{D}+L)^{-1} b \end{cases} \qquad (11)$$

Zauważmy, że zmieniając kolejność pętli w (9) na odwrotną (tj. for $i = n:1$ step -1 do) otrzymamy alternatywny wariant metody G-S o następującej postaci macierzowo-wektorowej.

$$\begin{cases} x^{k+1} = \tilde{G}_{gs} x^{k} + \tilde{W}_{gs} \\ \tilde{G}_{gs} = -(\mathcal{D}+U)^{-1} L, \quad \tilde{W}_{gs} = (\mathcal{D}+U)^{-1} b \end{cases} \qquad (12)$$

c) Metoda nadrelaksacji (SOR - Successive Over Relaxation)

Kolejnym ulepszeniem jest wprowadzenie nadrelaksacji. Punktem wyjścia jest metoda Gaussa-Seidela w formie zadanej wzorem (9). Metoda SOR przedstawia się następująco:

$$\text{for } i = 1:n \text{ do}$$
$$t_i = \frac{1}{a_{ii}}\left(b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{k+1} - \sum_{j=i+1}^{n} a_{ij} x_j^{k}\right) \qquad (13)$$
$$x_i^{k+1} = \omega t_i + (1-\omega) x_i^{k}$$
$$\text{end;}$$

W metodzie nadrelaksacji, parametr ω przyjmuje wartość większą od 1. Zauważmy, że dla $\omega = 1$ metoda SOR sprowadza się do metody G-S.

Celem stosowania nadrelaksacji jest przyspieszenie zbieżności procesu iteracyjnego. Wartość parametru ω wpływa na wielkość promienia spektralnego macierzy G_{SOR} odpowiadającej metodzie SOR - wybierając odpowiednio ω można ten promień zmniejszyć (w porównaniu z metodą G-S) i znacznie przyspieszyć zbieżność. Dla pewnych klas macierzy (spotykanych w zastosowaniach) istnieją twierdzenia o optymalnym wyborze parametru ω.

Zadanie: wyprowadzić macierzowo-wektorową postać metody SOR

$$x^{k+1} = G_{SOR} x^k + W_{SOR} \qquad (14)$$

Odp. $G_{SOR} = (D - \omega L)^{-1}[(1-\omega)D + \omega U]$, $W_{SOR} = \omega (D - \omega L)^{-1} b$.

4. Zbieżność klasycznych metod iteracyjnych

Ogólne kryterium zbieżności stacjonarnego procesu iteracyjnego $x^{k+1} = Gx^k + W$ wymaga, aby promień spektralny macierzy G był mniejszy od jedności. W praktyce, potrzebujemy kryteriów dotyczących bezpośrednio macierzy układu $Ax = b$, a nie macierzy procesu iteracyjnego. Innymi słowy, interesują nas warunki, które powinna spełniać macierz A taki, aby macierz G tej czy innej metody iteracyjnej miała promień spektralny mniejszy od 1.

Zacznijmy od zdefiniowania kilku własności.

DEF: Mówimy, że macierz kwadratowa A o wymiarze n jest:

1) diagonalnie dominująca jeżeli

$$|a_{ii}| \geq \sum_{j \neq i} |a_{ij}|, \quad \text{dla } i = 1, 2, .., n \qquad (15)$$

Jeżeli powyższa nierówność jest OSTRA dla każdej wartości indeksu i to macierz A jest SILNIE diagonalnie dominująca

2) redukowalna jeśli istnieje taka macierz permutacyjna P (jest to macierz jednostkowa o poprzestawianych kolumnach lub - co równoważne - wierszach), że:

$$PAP^T = \begin{bmatrix} B_{11} & B_{12} \\ 0 & B_{22} \end{bmatrix} \rightarrow \text{blok kwadratowy}$$

i nieredukowalna jeśli takiej macierzy P nie ma.

3) nieredukowalna i diagonalnie dominująca jeśli jest diagonalnie dominująca, nieredukowalna i przynajmniej dla jednej wartości „i" nierówność (15) jest ostra.

4) nieujemna (co piszemy $A \geq 0$) jeśli wszystkie elementy macierzy są nieujemne, i dodatnia (co piszemy $A > 0$) jeśli wszystkie elementy są dodatnie.

5) M-macierzą jeśli wszystkie elementy pozadiagonalne (a_{ij}, $i \neq j$) są niedodatnie i jednocześnie macierz odwrotna do A jest nieujemna ($A^{-1} \geq 0$).

Podamy bez dowodu kluczowe twierdzenia:

TW 1: Niech macierz A będzie macierzą nieredukowalną i diagonalnie dominującą lub silnie diagonalnie dominującą. Wówczas metody Jacobiego i Gaussa-Seidela są zbieżne do rozwiązania $Ax = b$ przy dowolnym wektorze startowym x^0.

TW 2: Załóżmy, że macierz A jest M-macierzą. Wówczas metody Jacobiego i Gaussa-Seidela są zbieżne do rozwiązania układu $Ax = b$ przy dowolnym wektorze startowym x^0

Dalej, mamy ważne twierdzenie o metodzie SOR.

TW 3 (Kahan): Załóżmy, że macierz A ma niezerowe elementy diagonalne ($a_{ii} \neq 0$, $i = 1, 2, \ldots, n$). Wówczas promień spektralny macierzy G_{SOR} spełnia nierówność

$$\rho(G_{SOR}) \geq |\omega - 1|.$$

Ponieważ warunkiem zbieżności jest, aby $\rho(G_{SOR}) < 1$ to musi zachodzić nierówność

$$|\omega - 1| < 1 \Rightarrow \omega \in (0, 2) \qquad (16)$$

Wniosek: Warunkiem koniecznym zbieżności metody SOR jest aby parametr ω należał do przedziału $(0,2)$ (w praktyce $\omega \in [1,2)$).

Ważnym przypadkiem układu liniowego jest przypadek, w którym macierz A jest (jednocześnie):

a) symetryczna ($A = A^T$, czyli $a_{ij} = a_{ji}$, $i,j = 1,2,..,n$)

b) dodatnio określona tj: dla dowolnego wektora $X \neq 0$ ma miejsce nierówność $x^T A x = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} x_i x_j > 0$.

Mamy dwa ważne twierdzenia:

TW4 (Ostrowskiego - Reicha): Jeżeli macierz A jest macierzą symetryczną i nieosobliwą (tj. $\det A \neq 0$), posiadającą dodatnie elementy diagonalne (tj. $a_{ii} > 0$, $i = 1,..,n$) oraz $\omega \in (0,2)$ to metoda iteracyjna SOR zbiega do rozwiązania układu liniowego $Ax = b$ przy dowolnym wektorze startowym X^0 wtedy i tylko wtedy, gdy macierz A jest dodatnio określona.

Wniosek: Dla macierzy symetrycznej i dodatnio określonej metoda Gaussa - Seidela jest zbieżna do rozwiązania przy dowolnym wektorze startowym X^0.

Ciekawe, że symetria i dodatnia określoność NIE WYSTARCZY do zapewnienia zbieżności metody Jacobiego. Dla tej metody mamy nieco „bardziej wymagające" twierdzenie

TW 5: Załóżmy, że macierz $A = D + B$ jest symetryczna (tj. $B = L + U = L + L^T$) i macierz diagonalna D zawiera wyłącznie liczby dodatnie ($a_{ii} > 0$, $i = 1, 2, \ldots, n$). Wówczas metoda iteracyjna Jacobiego zbiega do rozwiązania układu $Ax = b$ przy dowolnym wektorze startowym x^0 wtedy i tylko wtedy, gdy zarówno A jak i macierz $\tilde{A} = D - B$ są macierzami dodatnio określonymi.

5. UWAGI nt. kryterium zatrzymania iteracji

Każdy proces iteracyjny musi mieć określone kryterium zatrzymania (stopu). Naturalnym kryterium stopu wydaje się być wielkość normy wektora błędu (np. norma maximum czyli $\|e\| := \max_{i=1,\ldots,n} |e_i|$)

$$\|e^k\| = \|x^k - x_*\|$$

problem w tym, że oczywiście x_* nie jest a priori znane!

W praktyce badamy zmianę rozwiązania w kolejnych iteracjach i stosujemy kryterium $\|x^{k+1} - x^k\| \leq \varepsilon$ (ε - mała liczba, np. $\varepsilon = 10^{-6}$)

ewentualnie $\|x^{k+1} - x^k\| \leq \tilde{\varepsilon} \cdot \|x^k\|$

Oba powyższe kryteria mogą być zawodne - w przypadku powolnej zbieżności, zmiany rozwiązania z reguły są niewielkie. Z tego powodu, po stwierdzeniu spełnienia jednej z powyższych nierówności, warto jest sprawdzić ile wynosi skalowana wartość residuum układu, tj. $\frac{1}{\|b\|}\|b - Ax^k\|$.