Komputerowa analiza zagadnień różniczkowych

1. Układy równań liniowych

P. F. Góra

http://th-www.if.uj.edu.pl/zfs/gora/

semestr letni 2007/08

Podstawowe fakty

- Równanie

$$\mathbf{A}\mathbf{x} = \mathbf{b}\,, \quad \mathbf{x}, \mathbf{b} \in \mathbb{R}^N, \mathbf{A} \in \mathbb{R}^{N\times N} \tag{1}$$

  ma jednoznaczne rozwiązanie gdy $\det \mathbf{A} \neq 0$.

- W praktyce numerycznej *nigdy* nie rozwiązuje się układu (1) "metodą wyznacznikową", bo jest ona niesłychanie kosztowna.

- Koszt numeryczny metod dokładnych $\sim O(N^3)$ dla macierzy pełnych, (znacznie) mniej dla macierzy rzadkich.

- Należy unikać *jawnego* obliczania odwrotności $\mathbf{A}^{-1}$, tym bardziej, że jest to konieczność egzotyczna. Dlatego *zawsze* zapis typu $\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$ rozumieć będziemy w ten sposób, że $\mathbf{x}$ spełnia równanie (1).

Metody dokładne

- Eliminacja Gaussa, *koniecznie* z wyborem elementu podstawowego ("pivotingiem"), częściowym lub pełnym.
- Rozkład LU: $\mathbf{A} = \mathbf{LU}$, gdzie $\mathbf{L}$ jest trójkątna dolna, $\mathbf{U}$ trójkątna górna, *koniecznie* z częściowym wyborem elementu podstawowego (algorytm Crouta) — zalecana metoda dla niewielkich macierzy dobrze uwarunkowanych, bez jakichś szczególnych symetrii. Koszt numeryczny rozkładu LU wynosi $O(N^3)$.
- Jeśli mamy rozkład LU $\mathbf{A} = \mathbf{LU}$, równanie (1) rozwiązujemy jako

$$\mathbf{Uy} = \mathbf{b} \tag{2a}$$
$$\mathbf{Lx} = \mathbf{y} \tag{2b}$$

  Równanie (2a) rozwiązujemy zaczynając od ostatniego wiersza idąc do góry (backsubstitution), równanie (2b) rozwiązujemy zaczynając od pierwszego wiersza idąc w dół (forward substitution) — w ten sposób za każdym razem rozwiązujemy równanie z jedną niewiadomą.

Rozkład Cholesky'ego

Jeżeli macierz $\mathbf{A}$ jest symetryczna i dodatnio określona, zamiast rozkładu LU można (i należy) używać rozkładu Cholesky'ego: $\mathbf{A} = \mathbf{L}\mathbf{L}^T$, gdzie $\mathbf{L}$ jest trójkątna dolna. Koszt rozkładu Cholesky'ego jest mniej więcej o połowę mniejszy od kosztu rozkładu LU.

Wersja Gaxpy:

for $j = 1 : n$

if $j > 1$

$$A(j:n, j) = A(j:n, j) - A(j:n, 1:j-1)A(j, 1:j-1)^T$$

end

$$A(j:n, j) = A(j:n, j)/\sqrt{A(j, j)}$$

end

Wersja Outer product:

for $k = 1 : n$

$A(k,k) = \sqrt{A(k,k)}$

$A(k+1:n,k) = A(k+1:n,k)/A(k,k)$

for $j = k+1 : n$

$A(j:n,j) = A(j:n,j) - A(j:n,k)A(j,k)$

end

end

Uwagi:

- Przy rozkładzie można nadpisywać macierz
- Nie da się robić pivotingu
- Jeśli $\mathbf{A}$ pasmowa, także $\mathbf{L}$ jest pasmowa, o takiej samej szerokości pasma
- Jeśli *wewnątrz* pasma występują dziury w $\mathbf{A}$, w $\mathbf{L}$ mogą one być wypełnione.

Norma macierzy

Niech $||\cdot||$ oznacza normę w przestrzeni $\mathbb{R}^N$:

Definicja: Niech $\mathbf{A} \in \mathbb{R}^{N\times N}$ będzie operatorem liniowym nad przestrzenią $\mathbb{R}^N$. Wielkość

$$||\mathbf{A}|| = \sup_{||\mathbf{x}||\neq 0} \frac{||\mathbf{Ax}||}{||\mathbf{x}||} \tag{3a}$$

nazywam *normą macierzy* $\mathbf{A}$ *indukowaną przez normę wektorów*.

Po prawej stronie powyższego wyrażenia występuje norma wektorów; to tłumaczy zastosowanie przymiotnika “indukowana”.

Obserwacja. Widać, iż

$$\|\mathbf{A}\| = \sup_{\|\mathbf{x}\|=1} \|\mathbf{A}\mathbf{x}\| . \tag{3b}$$

Twierdzenie 1. *Norma indukowana jest normą w przestrzeni operatorów liniowych nad* $\mathbb{R}^N$.

Definicja. Niech $\mathbf{A} \in \mathbb{R}^{N\times N}$ i $\det \mathbf{A} \neq 0$. Wielkość

$$\kappa = \left\|\mathbf{A}^{-1}\right\| \left\|\mathbf{A}\right\| \tag{4}$$

nazywam *współczynnikiem uwarunkowania* macierzy $\mathbf{A}$. Jeżeli $\det \mathbf{A} = 0$, przyjmuję $\kappa = \infty$.

Twierdzenie 2. *Niech $\mathbf{A} \in \mathbb{R}^{N\times N}$ będzie macierzą symetryczną, $\mathbf{A} = \mathbf{A}^T$, i niech liczby $\{\lambda_i\}_{i=1}^N$ będą jej wartościami własnymi. Jeżeli $\det \mathbf{A} \neq 0$, współczynnik uwarunkowania tej macierzy spełnia*

$$\kappa = \frac{\max\limits_i |\lambda_i|}{\min\limits_i |\lambda_i|}\,. \tag{5}$$

Dowód. W celu udowodnienia tego twierdzenia obliczmy normę macierzy $\mathbf{A}$. Ponieważ jest to macierz symetryczna i rzeczywista, jej wartości własne są rzeczywiste, natomiast jej unormowane wektory własne tworzą bazę w $\mathbb{R}^N$. Oznaczmy przez $\mathbf{y}_i$ jej i-ty wektor własny, $\mathbf{A}\mathbf{y}_i = \lambda_i \mathbf{y}_i$. Każdy wektor $\mathbf{x} \in \mathbb{R}^N$, $\|\mathbf{x}\| = 1$, można przedstawić jako kombinację liniową

$$x = \sum_{i=1}^{N} \alpha_i \mathbf{y}_i \,, \tag{6}$$

przy czym warunek unormowania prowadzi do następujące więzu na współczynniki tej kombinacji:

$$\sum_{i=1}^{N} \alpha_i^2 = 1 \,. \tag{7}$$

Obliczmy teraz

$$\begin{aligned}
\|\mathbf{Ax}\|^2 &= \left\|\mathbf{A}\sum_{i=1}^{N}\alpha_i\mathbf{y}_i\right\|^2 = \left\|\sum_{i=1}^{N}\alpha_i\mathbf{A}\mathbf{y}_i\right\|^2 = \left\|\sum_{i=1}^{N}\alpha_i\lambda_i\mathbf{y}_i\right\|^2 \\
&\leqslant \sum_{i=1}^{N}||\alpha_i\lambda_i\mathbf{y}_i||^2 = \sum_{i=1}^{N}\alpha_i^2\lambda_i^2 \leqslant \sum_{i=1}^{N}\alpha_i^2 \max_i \lambda_i^2 \\
&= \max_i \lambda_i^2 \sum_{i=1}^{N}\alpha_i^2 = \left(\max_i |\lambda_i|\right)^2 \qquad (8)
\end{aligned}$$

Widzimy zatem, iż $\forall\, \mathbf{x} \in \mathbb{R}^N$, $\|\mathbf{x}\|^2 = 1$, zachodzi $\|\mathbf{A}\mathbf{x}\| \leqslant \max_i |\lambda_i|$, a zatem na mocy definicji (3) $\|\mathbf{A}\| = \max_i |\lambda_i|$.

Ponieważ $\det \mathbf{A} \neq 0$, macierz $\mathbf{A}^{-1}$ istnieje, jest symetryczna i rzeczywista, a jej wartościami własnymi są liczby $\{1/\lambda_i\}_{i=1}^N$. Zupełnie analogicznie dowodzimy, iż $\|\mathbf{A}^{-1}\| = \max_i (1/|\lambda_i|) = 1/\left(\min_i |\lambda_i|\right)$, skąd natychmiast wynika teza (5).

□

Singular Value Decomposition

Twierdzenie 3. *Dla każdej macierzy* $\mathbf{A} \in \mathbb{R}^{M \times N}$, $M \geqslant N$, *istnieje rozkład*

$$\mathbf{A} = \mathbf{U}\,[\mathrm{diag}(w_i)]\,\mathbf{V}^T, \tag{9}$$

gdzie $\mathbf{U} \in \mathbb{R}^{M \times N}$ *jest macierzą kolumnowo ortogonalną,* $\mathbf{V} \in \mathbb{R}^{N \times N}$ *jest macierzą ortogonalną oraz* $w_i \in \mathbb{R}$, $i = 1, \ldots, N$. *Rozkład ten nazywamy* rozkładem względem wartości osobliwych *(Singular Value Decomposition, SVD). Jeżeli* $M = N$, *macierz* $\mathbf{U}$ *jest macierzą ortogonalną.*

Jądro i zasięg operatora

Niech $\mathbf{A} \in \mathbb{R}^{M \times N}$. *Jądrem operatora* $\mathbf{A}$ nazywam

$$\text{Ker}\,\mathbf{A} = \{\mathbf{x} \in \mathbb{R}^N : \mathbf{A}\mathbf{x} = 0\}\,. \tag{10}$$

Zasięgiem operatora $\mathbf{A}$ nazywam

$$\text{Range}\,\mathbf{A} = \{\mathbf{y} \in \mathbb{R}^M : \exists \mathbf{x} \in \mathbb{R}^N : \mathbf{A}\mathbf{x} = \mathbf{y}\}\,. \tag{11}$$

Jądro i zasięg operatora są przestrzeniami liniowymi. Jeśli $M = N < \infty$, $\dim(\text{Ker}\,\mathbf{A}) + \dim(\text{Range}\,\mathbf{A}) = N$.

Sens *SVD*

Sens *SVD* najlepiej widać w przypadku, w którym co najmniej jedna z wartości $w_i = 0$. Dla ustalenia uwagi przyjmijmy $w_1 = 0$, $w_{i\neq 1} \neq 0$.
Po pierwsze, co to jest $\mathbf{z} = [z_1, z_2, \ldots, z_n]^T = \mathbf{V}^T\mathbf{x}$? Ponieważ $\mathbf{V}$ jest macierzą ortogonalną, $\mathbf{z}$ jest rozkładem wektora $\mathbf{x}$ w bazie kolumn macierzy $\mathbf{V}$. Korzystając z (9), dostajemy

$$\mathbf{A}\mathbf{x} = \mathbf{U}\left[\mathrm{diag}(w_i)\right]\mathbf{V}^T\mathbf{x} = \mathbf{U}\left[\mathrm{diag}(0, w_2, \ldots, w_N)\right]\mathbf{z} = \mathbf{U}\begin{bmatrix} 0 \\ w_2 z_2 \\ \vdots \\ w_N z_N \end{bmatrix}. \qquad (12)$$

Wynikiem ostatniego mnożenia będzie pewien wektor z przestrzeni $\mathbb{R}^M$. Ponieważ pierwszym elementem wektora $[0, w_2 z_2, \ldots, w_N z_N]^T$ jest zero, wynik ten nie zależy od pierwszej kolumny macierzy $\mathbf{U}$. Widzimy zatem, że *kolumny macierzy $\mathbf{U}$, odpowiadające niezerowym współczynnikom w_i, stanowią bazę w zasięgu operatora $\mathbf{A}$*.
Co by zaś się stało, gdyby $\mathbf{x}$ był równoległy do wektora stanowiącego pierwszą kolumnę $\mathbf{V}$? Wówczas $\mathbf{z} = 0$, a wobec tego $\mathbf{A}\mathbf{x} = 0$. Ostatecznie więc widzimy, że *kolumny macierzy $\mathbf{V}$, odpowiadające zerowym współczynnikom w_i, stanowią bazę w jądrze operatora $\mathbf{A}$.*

SVD i odwrotność macierzy

Niech $\mathbf{A} \in \mathbb{R}^{N\times N}$. Zauważmy, że $|\det \mathbf{A}| = \prod_{i=1}^{N} w_i$, a zatem $\det \mathbf{A} = 0$ wtedy i tylko wtedy, gdy co najmniej jeden $w_i = 0$. Niech $\det \mathbf{A} \neq 0$. Wówczas równanie $\mathbf{A}\mathbf{x} = \mathbf{b}$ ma rozwiązanie postaci

$$\mathbf{x} = \mathbf{A}^{-1}\mathbf{b} = \mathbf{V}\left[\mathrm{diag}(w_i^{-1})\right]\mathbf{U}^T\mathbf{b}\,. \tag{13}$$

Niech teraz $\det \mathbf{A} = 0$. Równanie $\mathbf{A}\mathbf{x} = \mathbf{b}$ *także* ma rozwiązanie, o ile tylko $\mathbf{b} \in \text{Range}\,\mathbf{A}$. Rozwiązanie to dane jest wzorem

$$\mathbf{x} = \tilde{\mathbf{A}}^{-1}\mathbf{b} = \mathbf{V}\left[\mathrm{diag}(\tilde{w}_i^{-1})\right]\mathbf{U}^T\mathbf{b}\,. \tag{14a}$$

gdzie

$$\tilde{w}_i^{-1} = \begin{cases} w_i^{-1} & \text{gdy } w_i \neq 0\,, \\ 0 & \text{gdy } w_i = 0\,. \end{cases} \tag{14b}$$

SVD i współczynnik uwarunkowania

Twierdzenie 4. *Jeżeli macierz* $\mathbf{A} \in \mathbb{R}^{N\times N}$ *posiada rozkład* (9) *oraz* $\det \mathbf{A} \neq 0$*, jej współczynnik uwarunkowania spełnia*

$$\kappa = \frac{\max\limits_i |w_i|}{\min\limits_i |w_i|}. \tag{15}$$

Jeśli macierz jest źle uwarunkowana, ale *formalnie* odwracalna, numeryczne rozwiązanie równania $\mathbf{Ax} = \mathbf{b}$ może być zdominowane przez wzmocniony błąd zaokrąglenia. Aby tego uniknąć, często zamiast (bezużytecznego!) rozwiązania dokładnego (13), używa się *przybliżonego* (i użytecznego!) rozwiązania w postaci (14) z następującą modyfikacją

$$\tilde{w}_i^{-1} = \begin{cases} w_i^{-1} & \text{gdy } |w_i| > \tau\,, \\ 0 & \text{gdy } |w_i| \leqslant \tau\,, \end{cases} \tag{16}$$

gdzie τ jest pewną zadaną tolerancją.

Metody iteracyjne

W metodach dokładnych otrzymane rozwiązanie jest dokładne z dokładnością do błędów zaokrąglenia, które, dodajmy, dla układów źle uwarunkowanych mogą być *znaczne*.

W metodach iteracyjnych rozwiązanie dokładne otrzymuje się, teoretycznie, w granicy nieskończenie wielu kroków — w praktyce liczymy na to, że po skończonej (i niewielkiej) ilości kroków zbliżymy się do wyniku ścisłego w granicach błędu zaokrąglenia.

Rozpatrzmy układ równań:

$$a_{11}x_1 + a_{12}x_2 + a_{13}x_3 = b_1 \quad (17a)$$
$$a_{21}x_1 + a_{22}x_2 + a_{23}x_3 = b_2 \quad (17b)$$
$$a_{31}x_1 + a_{32}x_2 + a_{33}x_3 = b_3 \quad (17c)$$

Przepiszmy ten układ w postaci

$$x_1 = (b_1 - a_{12}x_2 - a_{13}x_3)/a_{11} \quad (18a)$$
$$x_2 = (b_2 - a_{21}x_1 - a_{23}x_3)/a_{22} \quad (18b)$$
$$x_3 = (b_3 - a_{31}x_1 - a_{32}x_2)/a_{33} \quad (18c)$$

Gdyby po prawej stronie (18) były “stare” elementy x_j, a po lewej “nowe”, dostalibyśmy metodę iteracyjną

$$x_i^{(k+1)} = \left(b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(k)} - \sum_{j=i+1}^{N} a_{ij} x_j^{(k)} \right) \Big/ a_{ii} \tag{19}$$

Górny indeks $x^{(k)}$ oznacza, ze jest to przybliżenie w k-tym kroku. Jest to tak zwana *metoda Jacobiego*.

Zauważmy, że w metodzie (19) nie wykorzystuje się najnowszych przybliżeń: Powiedzmy, obliczając $x_2^{(k+1)}$ korzystamy z $x_1^{(k)}$, mimo iż znane jest już wówczas $x_1^{(k+1)}$. (Za to metodę tę łatwo można zrównoleglić.) Sugeruje to następujące ulepszenie:

$$x_i^{(k+1)} = \left(b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(k+1)} - \sum_{j=i+1}^{N} a_{ij} x_j^{(k)} \right) \Big/ a_{ii} \tag{20}$$

Jest to tak zwana metoda Gaussa-Seidela.

Jeżeli macierz $\mathbf{A} = \{a_{ij}\}$ jest rzadka, obie te metody iteracyjne będą efektywne *tylko i wyłącznie* wówczas, gdy we wzorach (19), (20) uwzględni się ich strukturę, to jest uniknie redundantnych mnożeń przez zera.

Trochę teorii

Metody Jacobiego i Gaussa-Seidela należą do ogólnej kategorii

$$\mathbf{M}\mathbf{x}^{(k+1)} = \mathbf{N}\mathbf{x}^{(k)} + \mathbf{b} \tag{21}$$

gdzie $\mathbf{A} = \mathbf{M} - \mathbf{N}$ jest *podziałem* (*splitting*) macierzy. Dla metody Jacobiego $\mathbf{M} = \mathbf{D}$ (część diagonalna), $\mathbf{N} = -(\mathbf{L} + \mathbf{U})$ (części pod- i ponaddiagonalne, bez przekątnej). Dla metody Gaussa-Seidela $\mathbf{M} = \mathbf{D} + \mathbf{L}$, $\mathbf{N} = -\mathbf{U}$. Rozwiązanie równania $\mathbf{A}\mathbf{x} = \mathbf{b}$ jest punktem stałym iteracji (21).

Definicja *Promieniem spektralnym* (diagonalizowalnej) macierzy $\mathbf{G}$ nazywam

$$\rho(\mathbf{G}) = \max\{|\lambda| : \exists \mathbf{y} \neq 0 : \mathbf{G}\mathbf{y} = \lambda\mathbf{y}\} \tag{22}$$

Twierdzenie 5. *Iteracja* (21) *jest zbieżna jeśli* $\det \mathbf{M} \neq 0$ *oraz* $\rho(\mathbf{M}^{-1}\mathbf{N}) < 1$.

Dowód. Przy tych założeniach iteracja (21) jest odwzorowaniem zwężającym. □

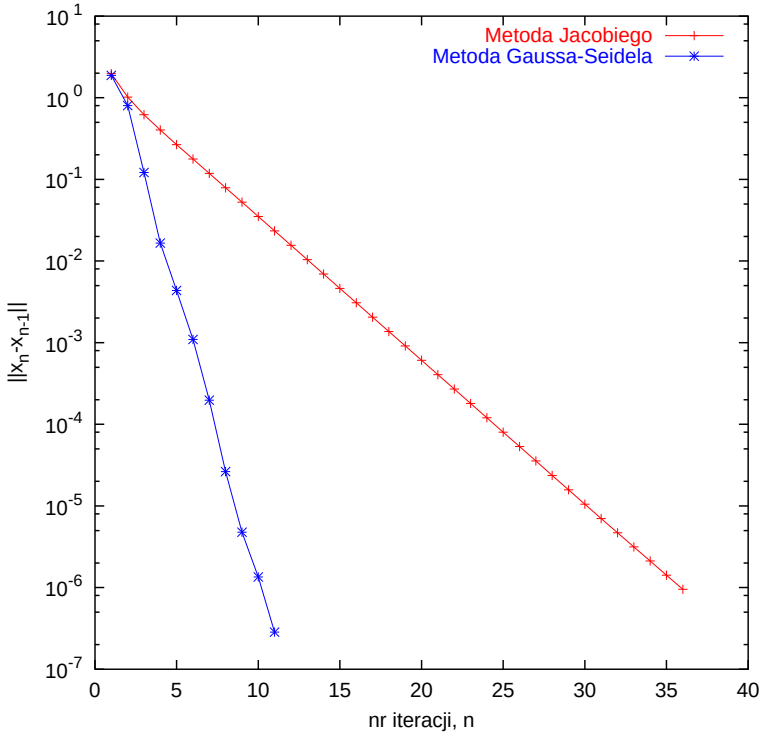
Twierdzenie 6. *Metoda Jacobiego jest zbieżna jeśli macierz* $\mathbf{A}$ *jest silnie diagonalnie dominująca.*

Twierdzenie 7. *Metoda Gaussa-Seidela jest zbieżna jeśli macierz* $\mathbf{A}$ *jest symetryczna i dodatnio określona.*

Przykład

Rozwiązujemy układ równań:

$$\begin{aligned} 3x &+ y &+ z &= 1 \\ x &+ 3y &+ z &= 1 \\ x &+ y &+ 3z &= 1 \end{aligned}$$



SOR

Jeśli $\rho(\mathbf{M}^{-1}\mathbf{N})$ w metodzie Gaussa-Seidela jest bliskie jedności, zbieżność metody jest bardzo wolna. Można próbować ją poprawić:

$$x_i^{(k+1)} = w\left(b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} - \sum_{j=i+1}^{N} a_{ij}x_j^{(k)}\right) \Big/ a_{ii} + (1-w)x_i^{(k)}, \tag{23}$$

gdzie $w \in \mathbb{R}$ jest *parametrem relaksacji*. Metoda ta zwana jest *succesive over-relaxation*, SOR. W postaci macierzowej

$$\mathbf{M}_w\mathbf{x}^{(k+1)} = \mathbf{N}_w\mathbf{x}^{(k)} + w\mathbf{b} \tag{24}$$

$\mathbf{M}_w = \mathbf{D} + w\mathbf{L}$, $\mathbf{N}_w = (1-w)\mathbf{D} - w\mathbf{U}$. *Teoretycznie* należy dobrać takie w, aby zminimalizować $\rho(\mathbf{M}_w^{-1}\mathbf{N}_w)$.