Metody numeryczne

Metody iteracyjne

Algebraiczna metoda gradientów sprzężonych

P. F. Góra

https://zfs.fais.uj.edu.pl/pawel_gora

4 listopada 2025

Metody iteracyjne

Rozwiązanie układu równań liniowych, uzyskane za pomocą którejś z dotąd poznanych metod, byłoby dokładne (ścisłe), gdyby nie błędy zaokrąglenia (które, dodajmy, dla układów źle uwarunkowanych mogą być *znaczne*). Dlatego metody te nazywa się *metodami dokładnymi*.

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

Metod iteracyjnych używa się *najczęściej*, choć nie wyłącznie, gdy zastosowanie faktoryzacji prowadziłoby do **wypełnienia** macierzy rzadkiej.

Rozpatrzmy układ równań:

$$a_{11}x_1 + a_{12}x_2 + a_{13}x_3 = b_1 \quad (1a)$$
$$a_{21}x_1 + a_{22}x_2 + a_{23}x_3 = b_2 \quad (1b)$$
$$a_{31}x_1 + a_{32}x_2 + a_{33}x_3 = b_3 \quad (1c)$$

Przepiszmy ten układ w postaci

$$x_1 = (b_1 - a_{12}x_2 - a_{13}x_3)/a_{11} \quad (2a)$$
$$x_2 = (b_2 - a_{21}x_1 - a_{23}x_3)/a_{22} \quad (2b)$$
$$x_3 = (b_3 - a_{31}x_1 - a_{32}x_2)/a_{33} \quad (2c)$$

Gdyby po prawej stronie (2) były “stare” elementy x_j, a po lewej “nowe”, dostalibyśmy metodę iteracyjną

$$x_i^{(k+1)} = \left(b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(k)} - \sum_{j=i+1}^{N} a_{ij} x_j^{(k)} \right) \Big/ a_{ii} \qquad (3)$$

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

W metodzie (3) nie wykorzystuje się najnowszych przybliżeń, **dzieki czemu metodę tę łatwo można zrównoleglić**. W metodzie Jacobiego obliczając $x_2^{(k+1)}$ korzystamy z $x_1^{(k)}$, mimo iż znane jest już wówczas $x_1^{(k+1)}$. Sugeruje to następujące ulepszenie:

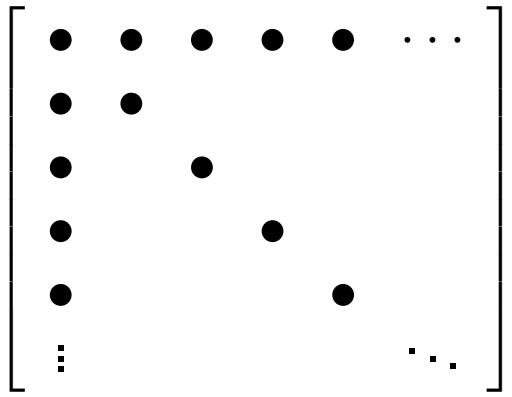
$$x_i^{(k+1)} = \left(b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(k+1)} - \sum_{j=i+1}^{N} a_{ij} x_j^{(k)} \right) \Big/ a_{ii} \qquad (4)$$

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 (3), (4) uwzględni się ich strukturę, to jest uniknie redundantnych mnożeń przez zera.

Powtórzmy: Dla numerycznej efektywności metod iteracyjnych jest **niesłychanie** ważne, aby metodę zaprogramować w ten sposób, aby uwzględniać strukturę macierzy rzadkiej.

Przykład: Niech macierz $\mathbf{A} \in \mathbb{R}^{N \times N}$ ma strukturę



(5)

Taka macierz jest rzadka, ma tylko $\sim 3N$ niezerowych elementów, domyślamy się więc, że układ równań z taką macierzą można rozwiązać w czasie liniowym. Zakładając, że macierz ta jest symetryczna i dodatnio określona, można by próbować zastosować do niej faktoryzację Cholesky'ego. Prowadziłoby to jednak do **wypełnienia** i okazałoby się, że cały algorytm "zyskałby" złożoność $O(N^3)$.

Metoda Gaussa-Seidela dla macierzy o strukturze (5) ma postać

$$
\begin{aligned}
x_1^{(k+1)} &= \left(b_1 - \sum_{j=2}^{N} a_{1j} x_j^{(k)}\right) / a_{11} \qquad (6)\\
x_2^{(k+1)} &= \left(b_2 - a_{21} x_1^{(k+1)}\right) / a_{22}\\
x_3^{(k+1)} &= \left(b_3 - a_{31} x_1^{(k+1)}\right) / a_{33}
\end{aligned}
$$

....

$$
x_N^{(k+1)} = \left(b_N - a_{N1} x_1^{(k+1)}\right) / a_{NN}
$$

Widać, że jedek krok (*sweep*) algorytmu (6) odbywa się w czasie proporcjonalnym do N.

Trochę teorii

Metody Jacobiego i Gaussa-Seidela należą do ogólnej kategorii

$$\mathbf{M}\mathbf{x}^{(k+1)} = \mathbf{N}\mathbf{x}^{(k)} + \mathbf{b} \tag{7}$$

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 (7).

Twierdzenie 1. *Iteracja* (7) *jest zbieżna jeśli* $\det \mathbf{M} \neq 0$ *oraz* $\rho(\mathbf{M}^{-1}\mathbf{N}) < 1$*, gdzie* $\rho(\bullet)$ *oznacza promień spektralny macierzy..*

Dowód. Przy tych założeniach iteracja (7) jest odwzorowaniem zwężającym. □

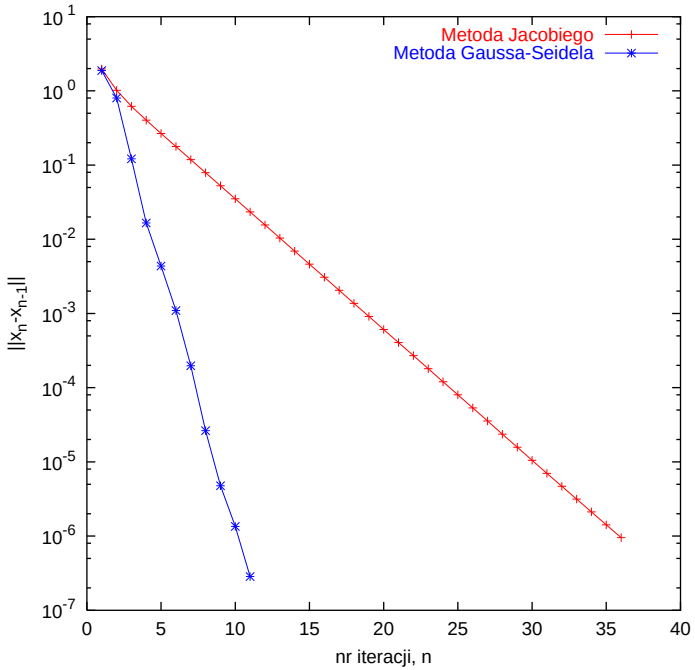
Twierdzenie 2. *Metoda Jacobiego jest zbieżna, jeśli macierz* $\mathbf{A}$ *jest silnie diagonalnie dominująca, to znaczy jeśli wartości bezwzględne elementów na głównej przekątnej są większe od sumy wartości bezwzględnych pozostałych elementów w danym wierszu.*

Twierdzenie 3. *Metoda Gaussa-Seidela jest zbieżna, jeśli macierz* $\mathbf{A}$ *jest symetryczna i dodatnio określona.*

Przykład

Rozwiązujemy układ równań:

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



Na maszynie jednoprocesorowej metoda Gaussa-Seidela jest szybsza, niż metoda Jacobiego. Gdybyśmy mogli zrównoleglić obliczenia, sytuacja mogłaby się odwrócić (gdyby obie metody były zbieżne).

Inny przykład

Dla macierzy o wymiarach 128×128

$$\begin{bmatrix} 128 & 1 & 1 & 1 & \dots & 1 \\ 1 & 2 & & & & \\ 1 & & 2 & & & \\ 1 & & & 2 & & \\ \vdots & & & & \ddots & \\ 1 & & & & & 2 \end{bmatrix} \tag{8}$$

(niezaznaczone elementy są zerami)

zbieżność z dokładnością do 10^{-12} w metodzie Gaussa-Seidela, według algorytmu (6), uzyskuje się w ~ 42 iteracjach.

Metoda gradientów sprzężonych — motywacja

Rozważmy funcję $f : \mathbb{R}^N \to \mathbb{R}$

$$f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x} - \mathbf{b}^T\mathbf{x} + c\,, \tag{9}$$

gdzie $\mathbf{x}, \mathbf{b} \in \mathbb{R}^N$, $c \in \mathbb{R}$, $\mathbf{A} = \mathbf{A}^T \in \mathbb{R}^{N\times N}$ jest *symetryczna i dodatnio określona*. Przy tych założeniach, funkcja (9) ma dokładnie jedno minimum, będące zarazem minimum globalnym. Szukanie minimów dodatnio określonych form kwadratowych jest (względnie) łatwe i z praktycznego punktu widzenia ważne. Minimum to leży w punkcie spełniającym

$$\nabla f = 0\,. \tag{10}$$

Obliczmy

$$\frac{\partial f}{\partial x_i} = \frac{1}{2}\frac{\partial}{\partial x_i}\sum_{j,k} A_{jk}x_j x_k - \frac{\partial}{\partial x_i}\sum_j b_j x_j + \underbrace{\frac{\partial c}{\partial x_i}}_{0}$$

$$= \frac{1}{2}\sum_{j,k} A_{jk}\left(\underbrace{\frac{\partial x_j}{\partial x_i}}_{\delta_{ij}} x_k + x_j \underbrace{\frac{\partial x_k}{\partial x_i}}_{\delta_{ik}}\right) - \sum_j b_j \underbrace{\frac{\partial x_j}{\partial x_i}}_{\delta_{ij}}$$

$$= \frac{1}{2}\sum_k A_{ik}x_k + \frac{1}{2}\sum_j A_{ji}x_j - b_i = \frac{1}{2}\sum_k A_{ik}x_k + \frac{1}{2}\sum_j A_{ij}x_j - b_i$$

$$= (\mathbf{Ax} - \mathbf{b})_i \ . \tag{11}$$

Widzimy zatem, że funkcja (9) osiąga minimum w punkcie, w którym zachodzi

$$\mathbf{A}\mathbf{x} - \mathbf{b} = 0 \Leftrightarrow \mathbf{A}\mathbf{x} = \mathbf{b}. \tag{12}$$

Rozwiązywanie układu równań liniowych (12) z macierzą symetryczną, dodatnio określoną jest równoważne poszukiwaniu minimum dodatnio określonej formy kwadratowej.

Przypuśćmy, że macierz $\mathbf{A}$ jest przy tym *rzadka* i duża (lub co najmniej średnio-duża). Wówczas metoda gradientów sprzężonych jest godną uwagi metodą rozwiązywania (12)

Metoda gradientów sprzężonych, *Conjugate Gradients*, CG

$\mathbf{A} \in \mathbb{R}^{N \times N}$ symetryczna, dodatnio określona, $\mathbf{x}_1$ — początkowe przybliżenie rozwiązania równania (12), $0 < \varepsilon \ll 1$.

$$
\begin{aligned}
&\mathbf{r}_1 = \mathbf{b} - \mathbf{A}\mathbf{x}_1,\ \mathbf{p}_1 = \mathbf{r}_1 \\
&\textbf{while } \|\mathbf{r}_k\| > \varepsilon \\
&\qquad \alpha_k = \frac{\mathbf{r}_k^T \mathbf{r}_k}{\mathbf{p}_k^T \mathbf{A}\mathbf{p}_k} \\
&\qquad \mathbf{r}_{k+1} = \mathbf{r}_k - \alpha_k \mathbf{A}\mathbf{p}_k \\
&\qquad \beta_k = \frac{\mathbf{r}_{k+1}^T \mathbf{r}_{k+1}}{\mathbf{r}_k^T \mathbf{r}_k} \\
&\qquad \mathbf{p}_{k+1} = \mathbf{r}_{k+1} + \beta_k \mathbf{p}_k \\
&\qquad \mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{p}_k \\
&\textbf{end}
\end{aligned}
\tag{13}
$$

Wówczas zachodzą twierdzenia:

Twierdzenie 4. *Ciągi wektorów* $\{\mathbf{r}_k\}$, $\{\mathbf{p}_k\}$ *spełniają następujące zależności:*

$$\mathbf{r}_i^T \mathbf{r}_j = 0\,, \quad i > j\,, \tag{14a}$$

$$\mathbf{r}_i^T \mathbf{p}_j = 0\,, \quad i > j\,, \tag{14b}$$

$$\mathbf{p}_i^T \mathbf{A} \mathbf{p}_j = 0\,, \quad i > j\,. \tag{14c}$$

Twierdzenie 5. *Jeżeli* $\mathbf{r}_M = 0$, *to* $\mathbf{x}_M$ *jest ścisłym rozwiązaniem równania* (12).

Dowód. Oba (sic!) dowody przebiegają indukcyjnie. □

Ciąg $\{\mathbf{x}_k\}$ jest w gruncie rzeczy "pomocniczy", nie bierze udziału w iteracjach, służy tylko do konstruowania kolejnych przybliżeń rozwiązania.

Istotą algorytmu jest konstruowanie dwu ciągów wektorów spełniających zależności (14). Wektory $\{\mathbf{r}_k\}$ są wzajemnie prostopadłe, a zatem *w arytmetyce dokładnej* $\mathbf{r}_{N+1} = 0$, wobec czego $\mathbf{x}_{N+1}$ jest poszukiwanym ścisłym rozwiązaniem.

Zauważmy, że ponieważ $\mathbf{A}$ jest symetryczna, dodatnio określona, warunek (14c) oznacza, że wektory $\{\mathbf{p}_k\}$ są wzajemnie prostopadłe w metryce zadanej przez $\mathbf{A}$. Ten właśnie warunek nazywa się warunkiem *sprzężenia względem* $\mathbf{A}$, co daje nazwę całej metodzie.

Ten wariant metody gradientów sprzężonych nazywamy "algebraicznym", gdyż przy założeniu, że *znamy* macierz $\mathbf{A}$ oraz wektor $\mathbf{x}_1$, możemy skonstruować ciągi $\{\mathbf{r}_k, \mathbf{p}_k, \mathbf{x}_k\}$ metodami algebraicznymi.

W przyszłości poznamy wariant metody gradientów sprzężonych, w którym wszystkich kroków nie uda się w ten sposób wykonać.

Koszt metody

W arytmetyce dokładnej metoda zbiega się po N krokach, zatem jej koszt wynosi $O(N \cdot \text{koszt\_jednego\_kroku})$. Koszt jednego kroku zdominowany jest przez obliczanie iloczynu $\mathbf{A}\mathbf{p}_k$. Jeśli macierz $\mathbf{A}$ jest pełna, jest to $O(N^2)$, a zatem całkowity koszt wynosi $O(N^3)$, czyli tyle, ile dla metod dokładnych. Jeżeli jednak $\mathbf{A}$ jest rzadka, koszt obliczania iloczynu jest mniejszy, **o ile obliczenie to jest odpowiednio zaprogramowane**. Jeśli $\mathbf{A}$ jest pasmowa o szerokości pasma $M \ll N$, całkowity koszt wynosi $O(M \cdot N^2)$.

Przykład

Dla macierzy o wymiarach 128×128

$$\begin{bmatrix} 128 & 1 & 1 & 1 & \dots & 1 \\ 1 & 2 & & & & \\ 1 & & 2 & & & \\ 1 & & & 2 & & \\ \vdots & & & & \ddots & \\ 1 & & & & & 2 \end{bmatrix} \tag{15}$$

(niezaznaczone elementy są zerami)

zbieżność z dokładnością do 10^{-12} w algebraicznej metodzie gradientów sprzężonych uzyskuje się po 4 (*sic!*) iteracjach (w metodzie Gaussa-Seidela były to 42 iteracje; w obu wypadkach znacznie poniżej rozmiaru macierzy).

Problem!

W arytmetyce o skończonej dokładności kolejne generowane wektory nie są *ściśle* ortogonalne do swoich poprzedników — na skutek akumulującego się błędu zaokrąglenia rzut na poprzednie wektory może stać się z czasem znaczny. Powoduje to istotne spowolnienie metody.

Twierdzenie 6. *Jeżeli* $\mathbf{x}$ *jest ścisłym rozwiązaniem równania* (12)*,* $\mathbf{x}_k$ *są generowane w metodzie gradientów sprzężonych, zachodzi*

$$\|\mathbf{x}-\mathbf{x}_k\| \leqslant 2\|\mathbf{x}-\mathbf{x}_1\| \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{k-1}, \tag{16}$$

gdzie κ *jest współczynnikiem uwarunkowania macierzy* $\mathbf{A}$*.*

Jeżeli $\kappa \gg 1$*, zbieżność może być bardzo wolna.*

Przykład

Rozwiązujemy układy równań z *małymi* (32×32) macierzami symetrycznymi, rzeczywistymi, dodatnio określonymi, o różnych współczynnikach uwarunkowania. Poniższy rysunek pokazuje normy kolejnych wektorów $\mathbf{r}_n$. Iteracje zatrzymywano, gdy $\|\mathbf{r}_n\| \leqslant 10^{-8}$. W arytmetyce dokładnej $\|\mathbf{r}_{n>32}\| \equiv 0$.