Wstęp do metod numerycznych

Zagadnienia wstępne

Uwarunkowanie

P. F. Góra

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

2018

Źródła błędów numerycznych

Wyniki obliczeń numerycznych obarczone są błędami. Ich najważniejszymi źródłami są

1. **Błędy grube**, pomyłki — zawinione przez człowieka. Czasami trudno je wyeliminować, ale teoretycznie wszystkie można usunąć, dlatego też, zalecając ostrożność i staranność autorom programów numerycznych, w analizie algorytmów pomijamy wpływ tych błędów.
2. **Błędy algorytmu**, wynikające z zastąpienia "idealnego" problemu matematycznego przez jakieś przybliżenie, na przykład zastąpienie zagadnienia obliczania całki oznaczonej przez obliczanie skończonego ciągu sum. W analizie numerycznej staramy się oszacować wpływ takich błędów. Algorytm, którego błędu nie umiemy oszacować, można uznać za bezwartościowy.

Dla niektórych problemów istnieją tak zwane algorytmy dokładne, które dałyby matematycznie ścisłe wyniki, gdyby można było pominąć...

3. Konsekwencje **błędów zaokrąglenia**, wynikających z tego, że nie wszystkie liczby dają się reprezentować na skończonych komputerach w sposób dokładny, a zatem obliczenia prowadzone są *ze skończoną precyzją (dokładnością)*. Tych błędów, poza bardzo szczególnymi przypadkami, nie daje się całkowicie wyeliminować, należy natomiast umieć szacować ich wpływ oraz go minimalizować.
   W ogólnym przypadku wpływ błędu algorytmu i błędu zaokrąglenia mogą się na siebie nakładać i wzajemnie wzmacniać.

Błąd zaokrąglenia

Sposoby reprezentacji liczb całkowitych i rzeczywistych — patrz wykład z Teoretycznych Podstaw Informatyki.

W sposób *ścisły* można reprezentować w komputerze tylko liczby całkowite (z pewnego zakresu) oraz liczby wymierne, posiadające *skończone* rozwinięcia binarne (z pewnego zakresu). Wszystkie inne liczby można reprezentować tylko w sposób przybliżony. Są one zatem obarczone pewnym błędem, zwanym *błędem zaokrąglenia*.

Rozwinięcia binarne

Niech liczba $x \in (0, 1)$. Wówczas jej rozwinięciem binarnym jest

$$x = \sum_{i=1}^{\infty} a_i \cdot 2^{-i} \tag{1}$$

gdzie współczynniki a_i mogą przybierać tylko wartości $\{0, 1\}$.

Przykład

Znajdźmy rozwinięcie binarne liczby $\frac{1}{5}$. Mamy

$$\begin{aligned}
\frac{1}{5} &= \frac{1}{8}+\frac{1}{5}-\frac{1}{8}=\frac{1}{8}+\frac{3}{40}=\frac{1}{8}+\frac{1}{8}\cdot\frac{3}{5} \\
&= \frac{1}{8}+\frac{1}{8}\left(\frac{1}{2}+\frac{3}{5}-\frac{1}{2}\right)=\frac{1}{8}+\frac{1}{16}+\frac{1}{8}\cdot\frac{6-5}{2\cdot 5}=\frac{1}{8}+\frac{1}{16}+\frac{1}{16}\cdot\frac{1}{5} \\
&= \frac{1}{8}+\frac{1}{16}+\frac{1}{16}\left(\frac{1}{8}+\frac{1}{16}+\frac{1}{16}\cdot\frac{1}{5}\right) \\
&= \frac{1}{8}+\frac{1}{16}+\frac{1}{128}+\frac{1}{256}+\frac{1}{256}\cdot\frac{1}{5} \\
&= \frac{1}{8}+\frac{1}{16}+\frac{1}{128}+\frac{1}{256}+\frac{1}{2048}+\frac{1}{4096}+\frac{1}{32768}+\frac{1}{65536} \\
&\quad +\frac{1}{65536}\cdot\frac{1}{5} \\
&= 2^{-3}+2^{-4}+2^{-7}+2^{-8}+2^{-11}+2^{-12}+2^{-15}+2^{-16}+\ldots \\
&= 0.001100110011(0011)_2 \qquad (2)
\end{aligned}$$

Otrzymaliśmy nieskończone, okresowe rozwinięcie binarne.

Analogicznie, ponieważ $\frac{1}{10} = \frac{1}{2} \cdot \frac{1}{5}$, a w rozwinięciach binarnych dzielenie przez 2 oznacza przesunięcie wszystkiego o jedną pozycję w prawo,

$$\frac{1}{10} = 0.0001100110011(0011)_2 \tag{3}$$

Widać stąd, że "naturalne" dla osób posługujących się systemem dziesiętnym ułamki $\frac{1}{10}$, $\frac{1}{100}$ itd *nie mogą być* ***dokładnie*** *reprezentowane* w pamięci. Komputer zapamiętuje tylko ich skończone przybliżenia, obarczone błędem zaokrąglenia. Po bardzo wielu krokach błąd ten może się skomasować do czegoś, co istotnie wpłynie na wynik obliczeń.

Zamiast więc iterować z krokiem $\frac{1}{10}$, iterujmy z krokiem $\frac{1}{8}$. Zamiast z $\frac{1}{100}$ weźmy $\frac{1}{128}$ itd.

Cyfry znaczące

Niech x będzie liczbą rzeczywistą, mającą ogólnie nieskończone rozwinięcie dziesiętne. Cyfry tego rozwinięcia numerujemy w sposób "naturalny": cyfra jedności ma numer zero, cyfra dziesiątek ma numer jeden, cyfra setek ma numer dwa itd. Cyfry części ułamkowej rozwinięcia dziesiętnego mają numery ujemne.

Liczba x jest poprawnie zaokrąglona na d-ej pozycji do liczby, którą oznaczamy $x^{(d)}$, jeśli błąd zaokrąglenia ε jest taki, że

$$|\varepsilon| = \left|x - x^{(d)}\right| \leqslant \frac{1}{2} \cdot 10^d \,. \tag{4}$$

Na przykład jeśli $x = 6.743996666\ldots$, $x^{(-3)} = 6.744$, $x^{(-7)} = 6.7439967$.

Jeśli $\bar{x}$ jest przybliżeniem dokładnej wartości x, to k-tą cyfrę dziesiętną liczby $\bar{x}$ nazwiemy *znaczącą*, jeśli

$$|x - \bar{x}| \leqslant \frac{1}{2} \cdot 10^k \tag{5}$$

oraz $|y| \geqslant 10^k$ (części ułamkowe rozinięcia mają numery ujemne!). Wynika stąd, że każda cyfra poprawnie zaokrąglonej liczby, począwszy od pierwszej cyfry różnej od zera, jest znacząca. Liczba cyfr znaczących jest pewną miarą błędu zaokrąglenia.

Propagacja błędu

Niech $\bar{x}$ będzie poprawnie zaokrąglonym do d przybliżeniem dokładnej liczby x. Można powiedzieć, że $x = \bar{x} + \varepsilon$, gdzie ε jest liczbą losową o rozkładzie jednostajnym w przedziale $\left[-\frac{1}{2}10^d, \frac{1}{2}10^d\right]$. Weźmy teraz dwie liczby $x = \bar{x} + \varepsilon_x$ oraz $y = \bar{y} + \varepsilon_y$. Co można powiedzieć o błędach powstałych w wyniku elementarnych obliczeń artymetycznych?

$$x + y = \bar{x} + \bar{y} + \underbrace{\varepsilon_x + \varepsilon_y}_{\varepsilon_+} . \tag{6}$$

Liczba ε_+ jest liczbą losową. Jeżeli obie liczby ε_x, ε_y mają rozkład jednostajny na przedziale $\left[-\frac{1}{2}10^d, \frac{1}{2}10^d\right]$, liczba ε_+ ma rozkład trójkątny na przedziale $\left[-10^d, 10^d\right]$.

Jeśli będziemy sumować $N >> 1$ liczb o błędach zaokrąglenia pochodzących z tego samego rozkładu, błąd sumy będzie dążył do rozkładu normalnego o szerokości proporcjonalnej do $\sqrt{N}\,10^d$.

W wypadku *mnożenia*,

$$x \cdot y \approx \bar{x} \cdot \bar{y} + \bar{x}\varepsilon_y + \bar{y}\varepsilon_x \,. \tag{7}$$

Jeżeli $|\bar{x}| \gg |\bar{y}|$ lub $|\bar{y}| \gg |\bar{x}|$, może to spowodować znaczny wzrost błędu (niewielki błąd multiplikuje się).

W wypadku *dzielenia*,

$$\frac{x}{y} \simeq \frac{\bar{x}}{\bar{y}} + \frac{1}{\bar{y}}\varepsilon_x + \frac{\bar{x}}{\bar{y}^2}\varepsilon_y \,. \tag{8}$$

Jeżeli $|\bar{y}| \ll |\bar{x}|$, błąd dzielenia może być bardzo duży! Dzielenie przez (względnie) małe liczby obarczone błędem, może powodować pojawienie się **znacznego** błędu ilorazu.

Wpływ błędów zaokrąglenia

Błędy zaokrąglenia — fakt, że na komputerach pracujemy w arytmetyce ze skończoną dokładnością — mają wpływ na prowadzone obliczenia. Wynik może zależeć od kolejności przeprowadzanych działań: dodawanie "na komputerze" nie jest łączne!

Przykład: Przypuśćmy, że prowadząc obliczenia z dokładnością do czterech cyfr znaczących chcemy znaleźć wartość sumy

$$1.000 + 0.0001 + 0.0001 + 0.0001 + 0.0001 + 0.0001 + 0.0001 + 0.0001 + 0.0001 + 0.0001 + 0.0001 \quad (9)$$

Zaczynamy sumować od lewej. Suma dwu pierwszych składników wynosi 1.0001. Ta liczba ma *pięć* cyfr znaczących, więc zostanie zaokrąglona do czterech cyfr znaczących. I tak okazuje się, że w przyjętej dokładności, $1.000 + 0.0001 = 1.000$. Widać zatem, że przy tym sposobie prowadzenia obliczeń, wartość *całej* sumy (9) wynosi 1.000.

Sumujmy teraz od prawej. $0.0001 + 0.0001 = 0.0002$. $0.0002 + 0.0001 = 0.0003$ i tak dalej. Suma dziesięciu pierwszych (od prawej) składników wynosi 0.0010. Gdy teraz dodamy tę wielkość do składnika ostatniego od prawej, otrzymamy 1.001. Ta liczba ma cztery cyfry znaczące, co mieści się w przyjętej dokładności i wyniku nie trzeba zaokrąglać.

Inny przykład

Rozważmy ciąg zadany przepisem

$$\begin{cases} x_{n+1} &= 4x_n - 1 \\ x_0 &= \frac{1}{3}\,. \end{cases} \tag{10}$$

Jak łatwo sprawdzić, ciąg (10) jest ciągiem stałym, którego wszystkie wyrazy są równe $\frac{1}{3}$. Co jednak się stanie, jeśli wyrazy tego ciągu będziemy *obliczali* posługując się arytmetyką przybliżoną, zachowując osiem cyfr znaczących?

Mamy zatem

$$x_0 = 0.33333333 \tag{11a}$$
$$4x_0 = 1.33333332 \tag{11b}$$

Ostatnia cyfra w powyższym wyrażeniu byłaby *dziewiątą* cyfrą znaczącą – nie mamy miejsca na jej przechowywanie, a więc musimy ją odrzucić. Zatem

$$4x_0 = 1.33333330 \tag{11c}$$
$$x_1 = 0.33333330 \tag{11d}$$

W podobny sposób wyliczamy $x_2 = 0.33333280$. Wyniki kolejnych iteracji przedstawia poniższa tabela:

n	x_n
0	0.33333333
1	0.3333333
2	0.3333328
3	0.3333312
4	0.3333248
5	0.3332992
6	0.3331968
7	0.3327872
8	0.3311488
9	0.3245952
10	0.2983808
11	0.1935232
12	−0.2259072
13	−1.9036288
14	−8.6145152
15	−35.458061
16	−142.83224

Jak widzimy, **w wyniku prowadzenia obliczeń ze skończoną precyzją**, ciąg stały zamienił się w ciąg monotonicznie rozbieżny do $-\infty$. Gdybyśmy

prowadzili obliczenia z większą – ale wciąż skończoną – precyzją, rezultat byłby taki sam, choć liczba stanów “przejściowych”, gdy wyrazy ciągu wciąż są bliskie ścisłej wartości $\frac{1}{3}$, zwiększyłaby się.

Uwarunkowanie zadania numerycznego

Niech $\varphi : \mathbb{R}^n \to \mathbb{R}^m$ będzie pewną funkcją odpowiednio wiele razy różniczkowalną i niech $\mathbf{x} \in \mathbb{R}^n$.

Definicja: Mówimy, że zagadnienie obliczenia $\varphi(\mathbf{x})$ jest *numerycznie dobrze uwarunkowane*, jeżeli niewielkie względne zmiany danych dają niewielkie względne zmiany rozwiązania. Zagadnienia, które nie są numerycznie dobrze uwarunkowane, nazywamy źle uwarunkowanymi.

Przykład

Rozważmy problem znalezienia rozwiązań równania

$$x^2 + bx + c = 0\,, \tag{12}$$

przy czym zakładamy, że $b^2 - 4c > 0$. Wiadomo, że rozwiązania mają w tym wypadku postać

$$x_{1,2} = \tfrac{1}{2}\left(-b \pm \sqrt{b^2 - 4c}\right)\,. \tag{13}$$

Jak dobrze uwarunkowane jest zagadnienie obliczania (13)? *Danymi* są tu współczynniki trójmianu, b, c. Zaburzmy te współczynniki: $b \to b + \varepsilon_2$, $c \to c + \varepsilon_3$.

Rozwiązaniami są teraz

$$\begin{aligned} \bar{x}_{1,2} &= \tfrac{1}{2}\left(-b+\varepsilon_2 \pm \sqrt{(b+\varepsilon_2)^2-4(c+\varepsilon_3)}\right) \\ &\simeq \tfrac{1}{2}\left(-b \pm \sqrt{b^2-4c}+\varepsilon_2 \pm \frac{2b\varepsilon_2-4\varepsilon_3}{2\sqrt{b^2-4c}}\right), \end{aligned} \tag{14}$$

gdzie dokonaliśmy rozwinięcia Taylora do pierwszego rzędu w $\varepsilon_{1,2}$. Widzimy, że błąd względny

$$\left|\frac{\bar{x}_{1,2}-x_{1,2}}{x_{1,2}}\right| \tag{15}$$

rośnie nieograniczenie, gdy $b^2-4c \to 0^+$. Problem wyznaczania pierwiastków trójmianu (12) jest wówczas numerycznie źle uwarunkowany. Problem ten jest dobrze uwarunkowany, gdy $b^2-4c \gg 0$.

Norma wektora

Niech $\mathcal{V}$ będzie pewną przestrzenią wektorową nad ciałem $\mathbb{C}$ (lub $\mathbb{R}$). *Normą wektora* nazywam funkcję $\|\cdot\| : \mathcal{V} \to \mathbb{R}$, spełniającą następujące warunki ($\mathbf{x}, \mathbf{y} \in \mathcal{V}$):

1. $\|\mathbf{x}\| \geqslant 0 \quad \wedge \quad \|\mathbf{x}\| = 0 \Leftrightarrow \mathbf{x} = 0.$
2. $\|\alpha\,\mathbf{x}\| = |\alpha| \cdot \|\mathbf{x}\| \quad \alpha \in \mathbb{C}.$
3. $\|\mathbf{x} + \mathbf{y}\| \leqslant \|\mathbf{x}\| + \|\mathbf{y}\|.$

Mówiąc niezbyt prezycyjnie, norma jest uogólnieniem pojęcia wartości bezwzględnej na przypadek wektorów.

Przykłady norm wektorów

W naszych rozważaniach przestrzeń liniowa $\mathcal{V}$ najczęściej będzie przestrzenią $\mathbb{R}^n$. Można w niej definiować wiele (różnych) norm. Najczęściej używa się jednej z trzech:

- Norma taksówkowa:

$$\|\mathbf{x}\|_1 = |x_1| + |x_2| + \cdots + |x_n| \tag{16a}$$

- Norma Euklidesowa:

$$\|\mathbf{x}\|_2 = \sqrt{\mathbf{x}^T\mathbf{x}} = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2} \tag{16b}$$

- Norma maximum (*worst offender*):

$$\|\mathbf{x}\|_\infty = \max_{i=1,\ldots,n} |x_i| \tag{16c}$$

Jeżeli nie zaznaczymy inaczej, przez normę wektorową będziemy rozumieć normę Euklidesową.

Współczynnik uwarunkowania

Niech $\varphi : \mathbb{R}^n \to \mathbb{R}^m$ będzie pewną funkcją, $\mathbf{x} \in \mathbb{R}^n$ dokładną wartością argumentu, a $\bar{\mathbf{x}} \in \mathbb{R}^n$ znanym numerycznym przybliżeniem $\mathbf{x}$.

Definicja: Jeżeli istnieje $\kappa \in \mathbb{R}$ taka, że

$$\forall \mathbf{x}, \bar{\mathbf{x}} : \quad \frac{\|\varphi(\mathbf{x}) - \varphi(\bar{\mathbf{x}})\|_{\mathbb{R}^m}}{\|\varphi(\mathbf{x})\|_{\mathbb{R}^m}} \leqslant \kappa \cdot \frac{\|\mathbf{x} - \bar{\mathbf{x}}\|_{\mathbb{R}^n}}{\|\mathbf{x}\|_{\mathbb{R}^n}} \tag{17}$$

nazywamy ją *współczynnikiem uwarunkowania* zagadnienia wyliczenia wartości $\varphi(\cdot)$ (względem zadanych norm).

Współczynnik uwarunkowania mówi jak bardzo błąd względny wyniku obliczeń "przekracza" błąd względny samej różnicy przybliżenia i wartości dokładnej. Spodziewamy się, że jeżeli przybliżenie znacznie różni się od wartości dokładnej, także wyniki obliczeń będą się znacznie różnić. W zagadnieniach numerycznie źle uwarunkowanych *może* się zdarzyć, że nawet niewielkie odchylenie przybliżenia od wartości dokładnej doprowadzi do znacznej różnicy wyników.

Układy równań liniowych

Niech $\mathbf{A} \in \mathbb{R}^{n\times n}$ będzie macierzą, $\mathbf{x}, \mathbf{b} \in \mathbb{R}^n$. Rozpatrujemy równanie

$$\mathbf{A}\mathbf{x} = \mathbf{b}\,, \tag{18}$$

Zakładamy, że macierz $\mathbf{A}$ oraz wektor wyrazów wolnych $\mathbf{b}$ są znane. Poszukujemy wektora $\mathbf{x}$. Równanie (18) jest równoważne następującemu układowi równań liniowych:

$$\left\{\begin{array}{ccccccccccc} a_{11}x_1 & + & a_{12}x_2 & + & a_{13}x_3 & + & \ldots & + & a_{1n}x_n & = & b_1 \\ a_{21}x_1 & + & a_{22}x_2 & + & a_{23}x_3 & + & \ldots & + & a_{2n}x_n & = & b_2 \\ a_{31}x_1 & + & a_{32}x_2 & + & a_{33}x_3 & + & \ldots & + & a_{3n}x_n & = & b_3 \\ \vdots & & \vdots & & \vdots & & \ddots & & \vdots & & \vdots \\ a_{n1}x_1 & + & a_{n2}x_2 & + & a_{n3}x_3 & + & \ldots & + & a_{nn}x_n & = & b_n \end{array}\right. \tag{19}$$

gdzie a_{ij} są elementami macierzy $\mathbf{A}$, natomiast x_j, b_j są elementami wektorów, odpowiednio, $\mathbf{x}, \mathbf{b}$.

Rozwiązywanie układów równan liniowych rzadko stanowi "samoistny" problem numeryczny. Zagadnienie to występuje jednak **bardzo często** jako pośredni etap wielu problemów obliczeniowych. Dlatego też dogłębna znajomość algorytmów numerycznego rozwiązywania układów równań liniowych jest niezwykle ważna.

Rozwiązywalność układów równań liniowych

Układ równań (18) ma jednoznaczne rozwiązanie wtedy i tylko wtedy, gdy

$$\det \mathbf{A} \neq 0 \,. \tag{20}$$

Z elementarnej algebry wiadomo, że rozwiązania można wówczas skonstruować posługując się *wzorami Cramera*. Uwaga: Numeryczne korzystanie ze wzorów Cramera jest **koszmarnie drogie** i dlatego w praktyce korzystamy z innych algorytmów.

Jak dobrze uwarunkowane jest zagadnienie rozwiązania równania (18)?

Przykład

Rozważmy następujące układy równań:

$$\begin{cases} 2x + 6y & = 8 \\ 2x + 6.00001y & = 8.00001 \end{cases} \qquad \begin{cases} 2x + 6y & = 8 \\ 2x + 5.99999y & = 8.00002 \end{cases}$$

Współczynniki tych układów równan różnią się co najwyżej o $0.00002 = 2 \cdot 10^{-5}$. Rozwiązaniem pierwszego są liczby $(1, 1)$, drugiego — liczby $(10, -2)$. Widzimy, że mała zmiana współczynników powoduje, że różnica rozwiązań jest $\sim 10^6$ razy większa, niż zaburzenie współczynników. Powyższe układy równań są źle uwarunkowane.

Norma macierzy

Niech $\mathbf{A} \in \mathbb{R}^{N \times N}$. *Normą macierzy* (indukowaną) nazywam

$$\|\mathbf{A}\| = \max\left\{\frac{\|\mathbf{A}\mathbf{x}\|}{\|\mathbf{x}\|} : \mathbf{x} \in \mathbb{R}^N, \mathbf{x} \neq 0\right\} = \max_{\|\mathbf{x}\|=1} \{\|\mathbf{A}\mathbf{x}\|\} \tag{21}$$

Promieniem spektralnym macierzy $\mathbf{A} \in \mathbb{R}^{N \times N}$ nazywam

$$\rho = \sqrt{\left\|\mathbf{A}\mathbf{A}^T\right\|} \tag{22}$$

Współczynnik uwarunkowania układu równań liniowych

Rozwiązujemy układ rownań ($\det \mathbf{A} \neq 0$)

$$\mathbf{A}\mathbf{y} = \mathbf{b} \tag{23a}$$

Przypuśćmy, że wyraz wolny $\mathbf{b}$ jest obarczony jakimś błędem $\boldsymbol{\Delta}\mathbf{b}$, czyli rozwiązujemy

$$\mathbf{A}\tilde{\mathbf{y}} = \mathbf{b} + \boldsymbol{\Delta}\mathbf{b} \tag{23b}$$

Zauważmy, że $\tilde{\mathbf{y}} - \mathbf{y} = \mathbf{A}^{-1}\left(\mathbf{b} + \boldsymbol{\Delta}\mathbf{b}\right) - \mathbf{A}^{-1}\mathbf{b} = \mathbf{A}^{-1}\boldsymbol{\Delta}\mathbf{b}$.

Jak błąd wyrazu wolnego wpływa na rozwiązanie? Obliczamy

$$\frac{\|\tilde{\mathbf{y}}-\mathbf{y}\|}{\|\mathbf{y}\|}=\frac{\left\|\mathbf{A}^{-1}\boldsymbol{\Delta}\mathbf{b}\right\|}{\|\mathbf{y}\|}\leqslant\frac{\left\|\mathbf{A}^{-1}\right\|\cdot\|\boldsymbol{\Delta}\mathbf{b}\|}{\|\mathbf{y}\|} \tag{24a}$$

Z drugiej strony

$$\|\mathbf{b}\|=\|\mathbf{A}\mathbf{y}\|\leqslant\|\mathbf{A}\|\cdot\|\mathbf{y}\|$$

skąd wynika, że

$$\frac{1}{\|\mathbf{y}\|}\leqslant\frac{\|\mathbf{A}\|}{\|\mathbf{b}\|} \tag{24b}$$

Ostatecznie

$$\frac{\|\tilde{\mathbf{y}}-\mathbf{y}\|}{\|\mathbf{y}\|}\leqslant\underbrace{\left\|\mathbf{A}\right\|\cdot\left\|\mathbf{A}^{-1}\right\|}_{\kappa}\cdot\frac{\|\boldsymbol{\Delta}\mathbf{b}\|}{\|\mathbf{b}\|} \tag{24c}$$

Współczynnik uwarunkowania macierzy symetrycznej, rzeczywistej

Niech $\mathbf{A} \in \mathbb{R}^{N\times N}$ będzie odwracalną macierzą symetryczną, rzeczywistą. W takim wypadku jej wartości własne są rzeczywiste a jej unormowane wektory własne $\{\mathbf{e}_i\}_{i=1}^N$ stanowią bazę ortogonalną w $\mathbb{R}^N$. Oznaczmy wartości własne tej macierzy przez $\{\lambda_i\}_{i=1}^N$. Weźmy dowolny $\mathbf{x} \in \mathbb{R}^N$ taki, że $\|\mathbf{x}\| = 1$. Wówczas

$$\mathbf{x} = \sum_{i=1}^N \alpha_i \mathbf{e}_i \,, \quad \sum_{i=1}^N \alpha_i^2 = 1 \,. \tag{25}$$

$$\begin{aligned} \|\mathbf{Ax}\| &= \left\| \mathbf{A} \sum_{i=1}^N \alpha_i \mathbf{e}_i \right\| = \left\| \sum_{i=1}^N \alpha_i \mathbf{A}\mathbf{e}_i \right\| = \left\| \sum_{i=1}^N \alpha_i \lambda_i \mathbf{e}_i \right\| = \sqrt{\sum_{i=1}^N \alpha_i^2 \lambda_i^2} \\ &\leqslant \sqrt{\sum_{i=1}^N \alpha_i^2 \max_j(\lambda_j^2)} = \max_j |\lambda_j| \cdot \sqrt{\sum_{i=1}^N \alpha_i^2} = \max_j |\lambda_j| \end{aligned} \tag{26}$$

Uwzględniając (21), widzimy, że $\|\mathbf{A}\| = \max_j |\lambda_j|$: norma odwracalnej macierzy symetrycznej, rzeczywistej jest równa największemu modułowi spośród jej wartości własnych.

Rozważmy teraz macierz $\mathbf{A}^{-1}$. Ma ona te same wektory własne, co $\mathbf{A}$, natomiast jej wartości własne są odwrotnościami wartości własnej macierzy nieodwróconej, $\mathbf{A}^{-1}\mathbf{e}_i = \frac{1}{\lambda_i}\mathbf{e}_i$. Postępując jak powyżej, łatwo możemy pokazać, że

$$\left\|\mathbf{A}^{-1}\right\| = \max_j \frac{1}{\left|\lambda_j\right|} = \frac{1}{\min_j \left|\lambda_j\right|}\,. \tag{27}$$

Widzimy zatem, że

Współczynnik uwarunkowania macierzy $\mathbf{A} \in \mathbb{R}^{n \times n}$ wynosi

$$\kappa = \|\mathbf{A}\| \cdot \left\|\mathbf{A}^{-1}\right\| \tag{28}$$

Dla macierzy symetrycznych, rzeczywistych sprowadza się to do

$$\kappa = \frac{\max\limits_i |\lambda_i|}{\min\limits_i |\lambda_i|}, \tag{29}$$

gdzie λ_i oznaczają wartości własne macierzy.